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

MELHORIA DE IMAGEM E VÍDEO SUBAQUÁTICOS POR APRENDIZAGEM AUTO-SUPERVISIONADA E TAREFA-PRETEXTO BASEADA EM DEGRADAÇÃO

por

Claudio Dornelles Mello Junior

Tese para obtenção do Título de
Doutor em Modelagem Computacional

Rio Grande, Janeiro, 2024

Claudio Dornelles Mello Junior

"MELHORIA DE IMAGEM E VÍDEO SUBAQUÁTICOS POR APRENDIZAGEM AUTO-SUPERVISIONADA E TAREFA-PRETEXTO BASEADA EM DEGRADAÇÃO"

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:

BANCA EXAMINADORA

ICP Brasil
Documento assinado digitalmente
SILVIA SILVA DA COSTA BOTELHO
Data: 17/01/2024 15:18:35-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Sílvia Silva da Costa Botelho(Orientadora - FURG)

Assinado de forma digital por Paulo Lilles Jorge Drews Junior:00124140009
Dados: 2024.01.25 12:14:30 -03'00'

Prof. Dr. Paulo Lilles Drews Jr (Coorientador - FURG)

gov.br
Documento assinado digitalmente
GLAUBER ACUNHA GONCALVES
Data: 24/01/2024 12:09:44-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Glauber Acunha Gonçalves (FURG)

gov.br
Documento assinado digitalmente
LEONARDO RAMOS EMMENDORFER
Data: 18/01/2024 10:34:55-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Leonardo Ramos Emmendorfer (UFSM)

gov.br
Documento assinado digitalmente
JOAO PAULO PAPA
Data: 26/01/2024 08:11:10-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. João Paulo Papa (UNESP)

gov.br
Documento assinado digitalmente
RICARDO MATSUMURA DE ARAUJO
Data: 21/01/2024 07:25:06-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Ricardo Matsumura Araújo (UFPel)

Rio Grande - RS
2024

Ficha Catalográfica

M527m Mello Junior, Claudio Dornelles.
Melhoria de imagem e vídeo subaquáticos por aprendizagem auto-supervisionada e tarefa-pretexto baseada em degradação / Claudio Dornelles Mello Junior. – 2024.
131 f.

Tese (doutorado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-Graduação em Modelagem Computacional, Rio Grande/RS, 2024.
Orientadora Dra. Sílvia Silva da Costa Botelho.
Coorientador: Dr. Paulo Lilles Drews Jr.

1. Realce 2. Imagem 3. Subaquático 4. Aprendizagem auto-supervisionada I. Botelho, Sílvia Silva da Costa II. Drews Jr., Paulo Lilles III. Título.

CDU 004

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

"Só se pode alcançar um grande êxito quando
nos mantemos fiéis a nós mesmos."
Friedrich Nietzsche

Para minha esposa Valeria e
meus filhos Gabriel e Bruno.

AGRADECIMENTOS

Este trabalho consolida as bases para uma nova linha de atuação profissional, que difere significativamente da minha formação técnica original. Sua concretização é o resultado de um grande esforço pessoal, intelectual, físico e familiar. Em Cada etapa finalizada múltiplos desafios foram atendidos a partir de infindáveis horas de estudo, pesquisa e programação. Entretanto, a conclusão destas etapas e do trabalho só foi possível com a contribuição de um conjunto de pessoas que tive (e tenho) a satisfação de conviver e com as quais tenho profunda gratidão e respeito.

Inicialmente, agradeço à minha orientadora, Profa. Dra. Sílvia Silva da Costa Botelho, e ao meu co-orientador, Prof. Dr. Paulo Lilles Jorge Drews Jr., pelo fundamental apoio no direcionamento e na condução das atividades, mas, principalmente, pelo conhecimento e experiência compartilhados, pela compreensão e respeito mostrados ao longo de todo o trabalho.

Agradeço também ao acadêmico de Engenharia de Computação Bryan Umpierre Moreira, pela parceria na execução das atividades, pelo interesse e pelo inestimável auxílio durante a coleta de dados e obtenção dos resultados.

Os momentos iniciais do trabalho foram particularmente difíceis, dada a profusão de novos conhecimentos que deveriam ser adquiridos. Neste período, destaco e agradeço a participação dos integrantes do projeto Netuno, em especial, dos colegas Dr. Cristiano Rafael Steffens, Dr. Matheus Machado do Santos e do engenheiro de automação Lucas Ricardo Vieira Messias, que demonstraram grande companheirismo, solidariedade e amizade, fatores que foram importantes na superação das dificuldades iniciais.

Agradeço à Universidade Federal do Rio Grande (FURG), através do Centro de Ciências Computacionais (C3) por me conceder a oportunidade de realizar este trabalho e, em especial, ao prof. Dr. Adriano Welasque Werhli pelo apoio na adequação das minhas atividades docentes durante a extenuante fase final do trabalho. Agradeço aos colegas do C3 pelo incentivo e aos professores do Programa de Pós-graduação em Modelagem Computacional (PPGMC) e do Programa de Pós-graduação em Computação (PPGComp) pelos ensinamentos, motivação e postura ética, dos quais tive a satisfação de ser aluno.

Agradeço especialmente ao meu pai, Claudio (*in memoriam*) pela particular visão de mundo, que hoje entendo e me identifico e, com grande carinho, à minha mãe, Lúcia, pelo exemplo de força e coragem, pelo incentivo permanente e pela fé e amor inabaláveis.

A realização deste trabalho se constituiu em uma longa e difícil jornada, compartilhada no dia-a-dia com minha família, minha esposa Valeria, acima de tudo, amiga e companheira, que me manteve na linha durante todo o caminho, e meus filhos Gabriel e Bruno, que representam a verdadeira motivação desta jornada. A vocês agradeço o amor, a parceria, o incentivo, o carinho e a compreensão nas ausências e nos momentos de isolamento.

RESUMO

Ao longo dos últimos anos, a melhoria de imagens subaquáticas têm configurado campo de investigação bastante dinâmico. Cenas submersas estão sujeitas aos efeitos da água, que provoca distorções na cor e contraste, perda na definição de bordas e contornos e *color cast*. A turbidez provocada por partículas em suspensão na água e a respectiva interação com a luz ambiente são as principais causas destas distorções. As propostas para a redução ou remoção destes efeitos concentram-se em métodos utilizando Processamento Digital de Imagens e métodos baseados em aprendizado profundo. O processamento digital de imagens têm mostrado resultados expressivos, a custa de uma rigidez metodológica. Em outro contexto, os métodos baseados em aprendizado profundo mostram-se progressivamente mais consistentes, com resultados bastantes efetivos. O presente trabalho apresenta proposta para realce de imagens e vídeos subaquáticos, baseado em aprendizado profundo e abordagem auto-supervisionada. Considera-se que uma imagem subaquática tende a apresentar distorção em algum nível e que esta distorção pode fornecer informações necessárias para a sua redução ou remoção. Na metodologia proposta, aumenta-se a degradação na imagem de saída de uma rede autoencoder de acordo com a presente na imagem de entrada. Gera-se, assim, uma imagem com degradação mais intensa, que é comparada a imagem de entrada na função de perda durante o treinamento. Este procedimento produz maior penalização na função de perda durante o treinamento, condicionando o aprendizado da rede a remover a distorção adicional. A imagem de entrada é, assim, decodificada na saída da rede com distorção reduzida. O algoritmo proposto não requer dados pareados ou processamento prévio, utilizando apenas como informação a imagem de entrada. A aplicação da metodologia proposta para realce de vídeo subaquático conta com estratégia para prover consistência temporal. Na proposta, considera-se que as *features* associadas ao ambiente não variam de forma significativa entre dois quadros sucessivos e que o resultado para a ação de realce deve ser o mesmo. O algoritmo realiza o ajuste nas intensidades dos pixels dos *frames*, evitando alterações intensas e repentinas no valor dos pixels na transição entre *frames*. Os resultados alcançados mostram que o método é efetivo tanto na recuperação de cor, contraste e textura como na redução na percepção da turbidez. Em relação a vídeo, além do realce, resultados obtidos com sequências reais mostram que a estratégia adotada provém consistência temporal.

Palavras-chaves: Realce, imagem, subaquático, aprendizagem auto-supervisionada.

ABSTRACT

In recent years, the underwater image quality improvement represents a very dynamic investigation area. Underwater scenes perception is affected by color and constrast distortion, blurring and color cast due to the water. The particles in suspension in the water and the respective interaction with ambient light are the main causes of these distortions. The proposals for reduction of these effects are related to digital image processing and deep learning approaches. Digital image processing methods show important results but with limited flexibility. In another context, the deep learning-based methods show a progressive consistency and effective results. In this work, a proposal for underwater image and video enhancement is presented. The method is based on neural networks and in self-supervised learning approach. The assumption considers that an underwater image presents degradation at some level and this degradation provides information that facilitate its reduction. The proposed algorithm increases the degradation on the decoded image of an autoencoder network. This additional degradation is in line with that shown by the input image. This degradation increased image and the input image are matched in loss function during the training stage. This procedure causes higher penalty in the loss function and the neural network learns to remove the degradation from decoded image. The algorithm requires no paired dataset or prior processing of the images. The input image is the only information required. In video enhancement is presented a time consistency methodology. The method considers that there are not significant changes in enhancement action between successive frames of the underwater video. The algorithm adjusts the pixel intensities of the frames avoiding the flickering effect caused by outlier values in pixel intensities that surge in the frame transitions. The results show the effectiveness of the method in turbidity reduction, and color, contrast and texture enhancement. Also, in video application, the adopted strategy for time consistency presents good performance on real underwater footages.

keywords: Underwater, enhancement, image, self-supervised learning

ÍNDICE

1 INTRODUÇÃO 19
1.1 Contextualização e justificativa 22
1.2 Hipóteses Científicas 26
1.3 Objetivos 26
2 FUNDAMENTAÇÃO TEÓRICA 28
2.1 Modelo físico da Formação da Imagem UW 28
2.2 Espaço de Cores 29
2.3 Autoencoder 30
2.4 Filtragem Gaussiana 32
2.5 Trabalhos Relacionados 33
3 ASPECTO CONCEITUAL DA PROPOSTA 50
3.1 Conceito 51
3.2 Função de degradação 54
4 METODOLOGIA PARA MELHORIA DA QUALIDADE DE IMAGENS SUBAQUÁTICAS 56
4.1 Módulo de Degradação 57
4.2 Mudanças no histograma 57
4.3 Ajustando cor e contraste 59
4.4 Preparando o realce de textura 63
4.4.1 Limitações ao tratamento de imagens 66
5 METODOLOGIA PARA MELHORIA DA QUALIDADE DE VÍDEO SUBAQUÁTICO 69
5.1 Desafios à melhoria de vídeo 69
5.2 Alterações no histograma e *flickering* 70
5.3 Máximos e mínimos como *outliers* 73
5.4 Módulo de degradação para realce de vídeo subaquático 75
5.4.1 Módulo de Geração dos Frames de Referência 77
5.4.2 Módulo de Geração dos Frames-Duais 79
6 METODOLOGIA DE IMPLEMENTAÇÃO 81
6.1 Rede Neural 81
6.2 *Dataset* UWData 84
6.3 Estratégias de treinamento 86
6.4 Função de perda 87
7 RESULTADOS 91
7.1 Resultados para o realce de imagens 92
7.1.1 Imagens de saída do modelo 92
7.1.2 Estudo comparativo 98
7.1.3 Resultados com o módulo de atenção para áreas saturadas 104
7.2 Resultados para o realce de vídeo 106

7.2.1 Estudo comparativo envolvendo métodos para realce subaquático . . . 110
7.2.2 Resultados para vídeos adquiridos em ambiente subaquático com reduzida iluminação e de alta turbidez 113
7.2.3 Limitações identificadas ao tratamento de vídeo 115
8 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS 117

LISTA DE FIGURAS

Figura 1.1: Efeitos sobre propagação da luz na água 19
Figura 1.2: Exemplos de imagens subaquáticas 20
Figura 2.1: Representação do espaço de cores RGB. Fonte: Burger and Burge (2016) . 30
Figura 2.2: Estrutura mínima de um autoencoder 31
Figura 2.3: Autoencoder com camadas ocultas paralelas 32
Figura 2.4: Filtro Gaussiano. (a) Característica 3D da distribuição do filtro gaussiano, (b) Matriz-*kernel* de ordem 5×5 e desvio-padrão 1. Fonte: Burger and Burge (2016) 32
Figura 2.5: Exemplos de imagens resultantes da filtragem gaussiana. (a) Imagem original, (b) Imagem após a filtragem gaussiana 33
Figura 3.1: Aspecto conceitual do método 51
Figura 4.1: Diagrama mostrando a estrutura do Módulo de Degradação 58
Figura 4.2: Imagem UW com suas versões com histograma expandido e contraído e respectivos histogramas. (a) Imagem original $I_{(k)}$, (b) Versão com expansão do histograma $I_{hst(k)}$ e (c) Versão com contração do histograma $I_{hsr(k)}$. 59
Figura 4.3: Exemplos de imagens subaquáticas mostrando cenas com diferentes níveis de degradação. (a) Imagem com baixo nível de degradação, (b) Imagem com degradação moderada-alta e *color cast* verde-azulado. 63
Figura 4.4: Imagem subaquática e versões resultantes das etapas de alteração no histograma e ajuste de cor e contraste. (a) Imagem de entrada $I_{(k)}$, (a1) Versão com histograma expandido $I_{hst(k)}$, (a2) Versão com reforço de cor e contraste $I_{ce(k)}$ (Eq. 4.7), (a3) Versão com histograma comprimido $I_{hsr(k)}$ e (a4) Versão com atenuação de cor e contraste $I_{cd(k)}$ (Eq. 4.6). (b) Histograma da imagem de entrada, (b1) Histograma de (a1), (b2) Histograma de (a2), (b3) Histograma de (a3), (b4) Histograma de (a4). . 64
Figura 4.5: Conjunto de imagens utilizadas pelo bloco de reforço de textura do MD. (a)Imagem de entrada $I_{(k)}$, (b) Imagem resultante do bloco de reforço de cor e contraste $I_{ce(k)}$, (c) Imagem resultante da filtragem gaussiana $\mathcal{G}(I_{ce(k)})$, (d) Imagem contendo apenas elementos de alta-frequência $I_{L(k)}$ e (e) Imagem guia para realce $I_{e(k)}$. 65
Figura 4.6: Conjunto de imagens utilizadas pelo bloco de atenuação de textura do MD. (a)Imagem de entrada $I_{(k)}$, (b) Imagem resultante do bloco de atenuação de cor e contraste $I_{cd(k)}$, (c) Imagem resultante da filtragem gaussiana $\mathcal{G}(I_{cd(k)})$, (d) Imagem com redução nos elementos de alta-frequência (numerador na Eq. 4.15) e (e) Imagem guia para degradação $I_{d(k)}$. 66

Figura 4.7: Exemplo de imagem de entrada, imagens-guia e respectivas distribuições no espaço de cores e NPD. (a)Imagem de entrada $I_{(k)}$, (b) Imagem $I_{d(k)}$, (c) Imagem $I_{e(k)}$ 66
Figura 4.8: Exemplo de imagem com acentuado desequilíbrio entre os canais de cores. (a) Imagem original, (b) Imagem com expansão do histograma e (c) Imagem realçada de saída. 67
Figura 5.1: Sequências de *frames* dos vídeos "Sonar"e "Mobula"e respectivas versões com histograma expandido. 71
Figura 5.2: Comportamento no tempo dos valores máximo, mínimo, médio e faixa dinâmica para os vídeos "Sonar", Ryan (2021) e "Mobula", Guerra (2022). 72
Figura 5.3: Comportamento no tempo dos valores médios e faixa dinâmica e taxa de variação da faixa dinâmica para o vídeo "Sonar". (a) Vídeo original, (b) Versão corrigida em 1% e (c) Versão corrigida em 3%. 75
Figura 5.4: Comportamento no tempo dos valores médios e faixa dinâmica e taxa de variação da faixa dinâmica para o vídeo "Mobula". (a) Vídeo original, (b) Versão corrigida em 1% e (c) Versão corrigida em 3%. 76
Figura 5.5: Diagrama do Módulo de Degradação para realce de vídeo 77
Figura 5.6: Diagrama do Módulo de Geração dos *Frames* de Referência (GFR) 78
Figura 6.1: Arquitetura do autoencoder implementado 82
Figura 6.2: Imagens de entrada e saída do autoencoder. (a) Entrada, (b) Saída. . . . 83
Figura 6.3: Característica de treinamento do autoencoder 83
Figura 6.4: Amostras de imagens do *dataset* UIEB 84
Figura 6.5: Amostras de imagens do *dataset* SUIM 85
Figura 6.6: Amostras de imagens do *dataset* RUIE 85
Figura 6.7: Amostras de imagens do *dataset*TURBID/Milk 86
Figura 6.8: Diagrama da representação da imagem interpretada como uma composição vetorial 88
Figura 7.1: Exemplo de imagens resultantes da função de degradação e respectivos histogramas e distribuição no espaço de cores RGB. (a) Imagem de entrada $I_{(k)}$, (b) Versão-guia para degradação $I_{d(k)}$, (c) Versão-guia para realce $I_{e(k)}$, (d) Imagem de saída do MD $\hat{I}^*_{(k)}$ e (e) Imagem de saída da rede $\hat{I}_{(k)}$. 93
Figura 7.2: Amostras de imagens de entrada e resultantes com e sem a presenças dos módulos para textura no modelo. (a) Imagem de entrada, (b) Saída sem módulos para textura, (c) Saída com módulos para textura. 94
Figura 7.3: Exemplos de imagens de entrada e resultantes da função de degradação e saída da rede. (a) Imagem de entrada $I_{(k)}$, (b) Imagem de saída da rede $\hat{I}_{(k)}$, (c) Imagem de saída do módulo de degradação $\hat{I}^*_{(k)}$. Imagens obtidas ao final do treinamento da rede. 95

Figura 7.4: Exemplos de imagens de entrada com objetos próximos à câmera e resultantes da função de degradação e saída da rede. (a) Imagem de entrada $I_{(k)}$, (b) Imagem de saída da rede $\hat{I}_{(k)}$, (c) Imagem de saída do módulo de degradação $\hat{I}^{*}_{(k)}$. Imagens obtidas ao final do treinamento da rede. 96
Figura 7.5: Exemplos de imagens de entrada com baixa turbidez. (a) Imagem de entrada, (b) Imagem de saída da rede. 97
Figura 7.6: Imagens das instalações da usina nuclear Fukushima Daiichi - Japão. (a) Imagem de entrada, (b) Imagem de saída da rede. 97
Figura 7.7: Imagens em ambiente atmosférico com cenas sujeitas à ação de neblina. (a) Imagem de entrada, (b) Imagem de saída da rede. 98
Figura 7.8: Exemplos de imagens resultantes dos métodos utilizados no estudo comparativo. 99
Figura 7.9: Imagem do *Turbid/Milk* com água sem turbidez significativa e a versão utilizada como imagem de referência no estudo comparativo. (a) Imagem Milk#100, (b) Distribuição dos pixels no espaço de cores de Milk#100, (c) Imagem utilizada como referência, e (d) Distribuição dos pixels no espaço de cores da imagem de referência. 103
Figura 7.10: Imagens resultantes dos métodos para imagens do *dataset* Turbid/Milk. . 103
Figura 7.11: Amostras de imagem com desequilíbrio entre os canais de cores atenuados pelo módulo de atenção. (a) Imagem original $I_{(k)}$, (b) Imagem realçada $\hat{I}_{(k)}$ sem o módulo de atenção no *pipeline*, (c) Imagem resultante do módulo de atenção $I_{at(k)}$ e (d) Imagem realçada $\hat{I}_{(k)}$ com o módulo de atenção no *pipeline*. 105
Figura 7.12: Frames de vídeos com respectivas versões realçadas. (a) Vídeo: "Octopus, (b) Vídeo: "Isla Socorro", (c) Vídeo: "Saint Martin" e (d) Vídeo: "Fish Farm". 109
Figura 7.13: Resultado comparativo para o vídeo "Los Arcos" (Detloff, 2022). Métodos: Saleh (Saleh et al., 2022), Islam (Islam et al., 2020), Mello2022 (Mello et al., 2022), Li (Li et al., 2020) e o proposto. 111
Figura 7.14: Resultado comparativo para o vídeo "Sonar" (Ryan, 2021). Métodos: Saleh (Saleh et al., 2022), Islam (Islam et al., 2020), Mello2022 (Mello et al., 2022), Li (Li et al., 2020) e o proposto. 112
Figura 7.15: Amostras de vídeos coletados em ambientes de alta degradação com respectivas versões realçadas. (a) Vídeo: EMA_TQ1 e (b) Vídeo: EMA_TQ2. 114

Figura 7.16: *Frames* #01 das sequências de EMA_TQ1 e EMA_TQ2 e respectivas distribuição no espaço de cores. (a) EMA_TQ1#1 e distribuição dos pixels, (b) Versão realçada de EMA_TQ1#1 e distribuição dos pixels (c) EMA_TQ2#1 e a distribuição dos pixels e (d) Versão realçada de EMA_TQ2#1 e a distribuição dos pixels. 116

LISTA DE TABELAS

Tabela 1: Classificação dos espaços de cores. Fonte: Ibraheem et al. (2012) 29
Tabela 2: Métodos de Restauração (Rest.)/Realce (Rce.) com abordagens não baseadas em DL. Abreviaturas: Objetivo (Obj.), Espaço de Cores (E.C.), Processamento Digital da Imagem (PDI). 45
Tabela 3: Métodos de Restauração (Rest.)/Realce (Rce.) baseados em DL. Abreviaturas: Objetivo (Obj.), Arquitetura da Rede (Arq. Rede), Aprendizado (Ap.), Espaço de Cores (E.C.), autoencoder (AE), Supervisionado (S), Não-Supervisionado (NS), *Ground-Truth* (GT). 46
Tabela 4: Métodos de Restauração (Rest.)/Realce (Rce.) baseados em DL - **Continuação**. Abreviaturas: Objetivo (Obj.), Arquitetura da Rede (Arq. Rede), Restauração (Rest.), Realce (Rce.), Aprendizado (Ap.), Espaço de Cores (E.C.), autoencoder (AE), Supervisionado (S), Não-Supervisionado (NS), *Ground-Truth* (GT). 47
Tabela 5: Métodos de Realce (Rce.) de vídeos **subaquáticos**. Abreviaturas: Objetivo (Obj.), Metodologia (Metod.), Realce (Rce.), *Dehazing* (Dhz.), *Denoising* (Dns.), Aprendizado (Ap.), Consistência temporal (C.T.), autoencoder (AE), Supervisionado (S), Não-Supervisionado (NS), *Ground-Truth* (GT). 48
Tabela 6: Resultados para as métricas UIQM, UCIQE e CCF. UDCP (Drews-Jr et al., 2016), Fusion (Ancuti et al., 2012), Mello2022 (Mello et al., 2022), Li2020 (Li et al., 2020), Li2021 (Li et al., 2021) e Islam (Islam et al., 2020).101
Tabela 7: Resultado quantitativo para as métricas com referência 104
Tabela 8: Resultado quantitativo para as métricas sem referência UIQM, UCIQE e CCF. Modelos sem e com o módulo de atenção são indicados por (Mod) e (Mod+Att), respectivamente. 106
Tabela 9: Resultado quantitativo para as métricas com imagem de referência PSNR, MSE, SSIM, CIEDE2000 (CIE2K), FSIM, FSIMc, and GSMD. Modelos sem e com o módulo de atenção são indicados por (Mod) e (Mod+Att), respectivamente. 106
Tabela 10: Resultados quantitativos para vídeo. 113

LISTA DE SÍMBOLOS

$\mathbf{A}$	Classe de funções de $\mathbb{G}^p$ para $\mathbb{F}^n$
A_g	Ganho associado a função gama
av_g	Média geral da imagem
$avg_{(k)}$	Média por canal de cor da imagem
$\mathbf{B}$	Classe de funções de $\mathbb{F}^n$ para $\mathbb{G}^p$
B^∞	Luz ambiente
b_w	Faixa dinâmica da imagem no espaço de cores
C	Componente da radiância da cena
c_1, c_2	Constantes de ponderação dos termos da função de perda
$C_{J(k)}$	Componente da imagem sem degradação na direção da imagem $I_{(k)}$
$C_{\hat{J}^*(k)}$	Componente da imagem degradada na direção da imagem
$C_{P(k)}$	Componente da luz ambiente da imagem sem degr. na direção da imagem $I_{(k)}$
$C_{\hat{P}^*(k)}$	Componente da luz ambiente da imagem degradada $\hat{I}^*_{(k)}$
d	Distância entre o observador e o objeto
E_c	Energia luminosa
F_D	*Decoder*
F_d	*Frame* guia para degradação
F_E	*Encoder*
F_e	*Frame* guia para realce
F_{cd}	*Frame* resultante do bloco de atenuação de cor e contraste
F_{ce}	*Frame* resultante do bloco de reforço de cor e contraste
F_{dgf}	*Frame* de referência para degradação
F_{egf}	*Frame* de referência para realce
F_{hsr}	*Frame* com histograma contraído
F_{hst}	*Frame* com histograma expandido
F_r	*Frame* de referência
$\mathcal{G}(\cdot)$	Filtragem gaussiana
gd	Fator atenuação-distância descrito no espaço de cores
g_r	Fator de reforço
H_{KG}	Kernel gaussiano
I	Imagem de entrada da rede
$\hat{I}$	Imagem de saída da rede
I^*	Imagem degradada
$\hat{I}^*$	Imagem de saída da função de degradação
I_{at}	Imagem resultante do módulo de atenção
I_d	Imagem guia para degradação do módulo de degradação
I_e	Imagem guia para realce do módulo de degradação

LISTA DE SÍMBOLOS (Continuação)

I_{cd}	Imagem de saída do bloco de atenuação de cor e contraste do MD
I_{ce}	Imagem de saída do bloco de reforço de cor e contraste do MD
I_L	Imagem contendo elementos de alta-frequência
I_x	Imagem com pixels de alta intensidade
J	Imagem associada à cena sem distorção provocada pela água
$\mathcal{K}$	Sistema de medida, Autoencoder
$\mathcal{L}$	Função de perda
max	Valor máximo
min	Valor mínimo
N	Número de termos
r_{bw}	Relação entre as seções não-usada e usada da faixa dinâmica
t	Fator de transmissão
thr	Valor limiar para um pixel ser considerado *outlier*
$\mathbf{X}$	Vetor de entrada do autoencoder
$\mathbf{Y}$	Vetor de saída do autoencoder
x	Informação a ser medida
y	Informação com ruído

Símbolos gregos

α	Ângulo entre a imagem sem degradação e a imagem com degradação
β	Ângulo entre a luz ambiente e a imagem com degradação
β_1	Decaimento exponencial para o primeiro momento (ADAM)
β_2	Decaimento exponencial para o segundo momento (ADAM)
Δ	Degradação, Distorção
Δ_f	Dissimilaridade ou distorção definida em $\mathbb{F}^n$
γ_k	Fator de atenuação da água por canal de cor
ϵ	Ruído presente na medida
ϵ_z	Constante de pequeno valor para estabilidade numérica (ADAM)
Λ_k	Parâmetro análogo à luz ambiente e definida no espaço de cor
σ	Desvio padrão
θ_D	Parâmetros do *decoder*
θ_E	Parâmetros do *encoder*

Sub índices

b	*backscattering*
bw	relativo à faixa dinâmica
c	câmera, cena, sem distorção
D	*decoder*
d	degradação
E	*encoder*
e	realce
f	*forward scattering*
hsr	relativo à contração do histograma
hst	relativo à expansão do histograma
im	imagem
J	referente à radiância da cena
k	canal de cor no espaço RGB
P	referente à luz ambiente que alcança o observador

LISTA DE ABREVIATURAS

ADAM	*Adaptive Moment Estimation*
AE	autoencoder
BRISQUE	*Blind/Referenceless Image Spatial Quality Evaluator*
CLAHE	*Contrast Limited Adaptative Histogram Equalization*
CNR	*Contrast to Noise Ratio*
DCP	*Dark Channel Prior*
DCT	*Discrete Cosine Transform*
DL	*Deep Learning*
DSNMF	*Deep Sparse Non-negative Matrix Factorization*
DSSIM	*Structural Dissimilarity Image Metric*
EMEE	*Enhancement Measure, Entropy*
FADE	*Fog Aware Density Evaluator*
FSIM	*Feature Similarity*
FSIMc	*Feature Similarity with Chrominance*
GAN	*Generative Adversarial Network*
GDL	*Gradient Difference Loss*
GFD	Módulo de Geração de *Frames*-guias Duais
GFR	Módulo de Geração de *Frames* de Referência
GMSD	*Gradient Magnitude Similarity Deviation*
HICRD	*Heron Island Coral Reef Dataset*
LSTM	*Long-Short Term Memory*
MD	Módulo de Degradação
MSE	*Mean Square Error*
NIQE	*Natural Image Quality Evaluator*
NIQEv	*Natural Image Quality Evaluator* aplicada a vídeo
PCQI	*Patch-based Contrast Quality Index*
PDI	Processamento Digital da Imagem
PSNR	*Peak Signal to Noise Ratio*
RAHIM	*Recursive Adaptative Histogram Modification*
ReLU	*Rectified Linear Unit*
RGHS	*Relative Global Histogram Stretching*
RMSE	*Root Mean Square Error*
RUIE	*Real Wolrd Underwater Image Enhancement*
SFM	*Structure From Motion*
SSL	*Self-Supervised Learning*
SUIM	*Semantic Segmentation Underwater Imagery*

LISTA DE ABREVIATURAS (Continuação)

UCCS	*Underwater Color Cast Set*
UCIQE	*Underwater Color Image Quality Evaluation*
UDCP	*Underwater Dark Channel Prior*
UIEBD	*Underwater Image Enhancement Benchmark Dataset*
UIQM	*Underwater Image Quality Metric*
UIQS	*Underwater Image Quality Set*
ULAP	*Underwater Light Attenuation Prior*
UW	*Underwater*

1 INTRODUÇÃO

Entre as diversas áreas do conhecimento associadas ao desenvolvimento dos recentes sistemas autônomos, a visão computacional destaca-se na perspectiva de interação homem-máquina e da atividade remota assistida ou automatizada. Operações subaquáticas (*Underwater* - UW) modernas, utilizam estes equipamentos e veículos em missões de manutenção, inspeção, arqueologia marinha, explorações ambientais, visando fauna e flora marinhas entre outras, Donaldson et al. (2020), Cong et al. (2021). A aquisição de imagens e sequências de vídeo durante a realização destas atividades são essenciais para o registro dos procedimentos e resultados.

A qualidade das imagens e vídeos subaquáticos é, essencialmente, afetada pelas condições físicas do ambiente onde a cena é observada. A profundidade da cena, o nível de iluminação disponível e a turbidez, caracterizada pelo tipo, deslocamento e quantidade do material em suspensão na água, são os principais fatores que degradam a qualidade. A densidade da água é cerca de 800 vezes maior que a do ar, Fayaz et al. (2020), e a luz visível é sensivelmente atenuada com o aumento da profundidade. As componentes da luz com grandes comprimentos de onda são mais fortemente atenuadas. Por exemplo, considerando a água com reduzida quantidade de materiais em suspensão (águas claras) e iluminação natural (solar) à superfície, a cor vermelha é quase que totalmente atenuada a partir dos 5m de profundidade, seguidas pelo laranja e o amarelo. O verde desaparece em torno dos 30m e o azul em 50m, Fayaz et al. (2020).

A água absorve a energia luminosa dos raios solares a medida que a distância percorrida aumenta (vertical e horizontalmente) e as partículas de material em suspensão provocam o desvio destes raios ao longo do seu percurso como representado na Fig.1.1.

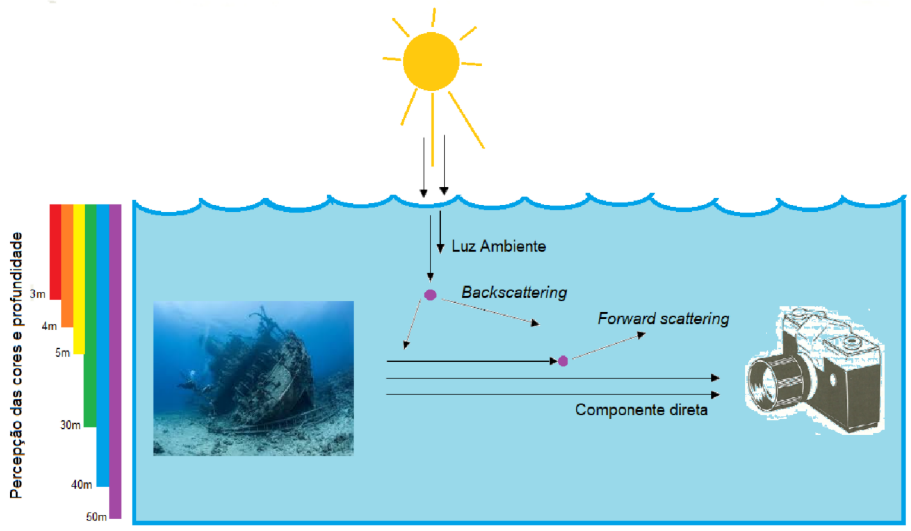


Figura 1.1: Efeitos sobre propagação da luz na água

Neste cenário, o sensor de uma câmera submersa percebe o feixe luz refletido por um objeto a uma dada distância (refletância). No trajeto até a câmera, alguns raios de luz são desviados pela colisão com as partículas em suspensão, dando origem ao fenômeno de espalhamento dos raios. Este efeito é identificado como *forward scattering* e reduz a informação dos objetos na

cena (componente direta) que incide na câmera. Isto causa borramento (*blurring*), com perda da percepção de detalhes da textura e nitidez da cena. Degradação adicional é ocasionada pela luz ambiente. Especificamente, os raios de luz do sol colidem com partículas do material em suspensão na água e se dispersam no ambiente. Uma parcela destes raios é desviada em direção à câmera, concorrendo com a irradiação da luz vinda dos objetos submersos. Este fenômeno é denominado *backscattering* e provoca aumento na percepção da turbidez, perda de contraste e borramento. Ambos os efeitos serão abordados em maior profundidade e à luz do modelo físico de formação da imagem, adiante neste texto.

Outro fator de degradação é a tonalidade azul, verde ou marrom que a água pode apresentar em função da natureza e quantidade do material em suspensão, Jerlov (1976). Este efeito é denominado de *color cast* e interfere na percepção de cor e contraste.

A percepção da intensidade de atuação destes fenômenos depende da natureza e quantidade de material em suspensão. Imagens (e vídeos) contendo cenas submersas tendem a apresentar algum nível de degradação. Observa-se que, mesmo em imagens adquiridas próximas a superfície da água e aparentando nível de turbidez muito reduzido, há atenuação na cor e perda de contraste a medida em que os objetos observados se distanciam da câmera.



Figura 1.2: Exemplos de imagens subaquáticas

Na Fig.1.2 são mostrados exemplos de imagens subaquáticas com indicação da degradação presente na cena. É notória a diversidade de situações impostas pela iluminação ambiente, turbidez e *color cast* na percepção de degradação na imagem. Esta diversidade apresenta o panorama do desafio proposto à melhoria na qualidade da imagem e vídeo subaquáticos. A real percepção de contraste, textura, intensidade e tonalidade das cores, ou seja, o que seria

percebido na cena da imagem sem a presença da água, não são conhecidas. Turbidez e *color cast* são os principais fatores que comprometem esta percepção. Esta condição, fundamentalmente, caracteriza a melhoria de imagem (e vídeo) submersa como um problema mal-posto. Rittel and Webber (1973) destacam que não existe uma formulação definitiva para um problema mal-posto e a busca da solução coincide com o entendimento de sua natureza. Um arcabouço metodológico concebido para melhoria da qualidade de imagens submersas requer a compreensão deste aspecto. E em termos concretos, deve atuar na redução de aspectos agnósticos da degradação como turbidez e o efeito da profundidade, assim como, lidar com a diversidade de contextos produzidos pela coloração dominante da água (*color cast*) e intensidade e distribuição da luz ambiente.

Ao longo dos últimos anos, a investigação científica relacionada à melhoria de imagens subaquáticas delineou os contextos de restauração e realce, Wang et al. (2019). Métodos de restauração fazem uso do modelo físico de formação da imagem na busca para compensar os efeitos da degradação impostos pela água, com destaque na remoção da turbidez. A estimativa de parâmetros físicos, pré-processamento das imagens ou uso de estruturas pré-treinadas caracterizam estas metodologias (Drews-Jr et al. (2016); Song et al. (2018); Anwar et al. (2020); Lin et al. (2020)). Nos métodos de realce, o foco reside na recuperação de cor e de contraste para melhoria da percepção humana da imagem, Li et al. (2020). A abordagem é orientada a readequação da intensidade dos pixels, utilizando processamento digital da imagem e/ou aprendizado profundo (*Deep Learning* - DL) e ocorre, fundamentalmente, no espaço de cores (Chiang and Chen (2012); Ancuti et al. (2018); Fabbri et al. (2018); Li et al. (2020)). Entretanto, resultados recentes apresentados pelas metodologias de realce mostram efetiva redução no efeito da turbidez e recuperação da percepção de bordas, antes associadas à propostas de restauração (Li et al. (2020); Mello et al. (2022); Saleh et al. (2022); Li et al. (2022a)).

O tratamento de imagens subaquáticas mostra evolução e ampla diversidade de metodologias, consolidadas na bibliografia como discutido acima. Entretanto, situação diversa é verificada no âmbito da melhoria da qualidade de vídeos UW. O número estudos e propostas mostra que esta linha de pesquisa ainda é pouco explorada. A melhoria da qualidade de vídeo é, consistentemente, mais desafiadora. Onde, além dos requisitos a serem atendidos no caso de imagem, exige que o resultado do método para melhoria da qualidade seja coerente no tempo. Isto é, a ação deve mostrar o mesmo resultado para cor, contraste e percepção da luz ambiente, mantendo coerência na intensidade e uniformidade em *frames* sucessivos de uma sequência de vídeo. Variações nesta ação de um *frame* para outro são facilmente percebidas em vídeos subaquáticos, identificando falta de efetividade do método de melhoria na questão temporal. A consistência temporal é linha de estudo consolidada em trabalhos relacionados com sequências em ambientes atmosféricos (*indoor, outdoor*), onde, por exemplo, a recuperação de imagens ou regiões de imagens sub-expostas (*low light*) é tema consistentemente investigado, Li et al. (2022b).

1.1 Contextualização e justificativa

Neste trabalho é apresentada proposta para melhoria de imagem e vídeo subaquáticos, fundamentada na aplicação conjunta de aprendizado profundo e elementos de processamento digital. Devido à natureza e à variabilidade dos fenômenos envolvidos, a tarefa enseja considerável grau de complexidade, particularmente, ao se considerar a capacidade de generalização que um modelo deve apresentar, dado a miríade de contextos encontrada no ambiente subaquático.

Modelos *deep learning* e a dependência dos dados

A investigação científica relacionada com realce UW estruturada em *deep learning* tem mostrado efetiva evolução ao longo dos últimos anos, comprovada através da quantidade e qualidade dos trabalhos apresentados, Raveendran et al. (2021). Nos *frameworks* propostos, destaca-se o emprego de redes convolucionais *end-to-end*, com destaque para redes *encoder-decoder* (Hashisho et al. (2019); Li et al. (2021); Yang et al. (2021)) e as *Generative Adversarial Networks* (GANs), utilizadas sobretudo em abordagens empregando *transfer learning*, (Fabbri et al. (2018); Ding et al. (2019); Cho et al. (2020)) ou *style transfer*, Yu et al. (2019).

Uma característica importante destes métodos é a dependência de dados anotados para o treinamento dos modelos. A construção de *datasets* pareados de imagens subaquáticas é bastante complexa, devido à dificuldade na captura de cenas livres da degradação imposta pela água. Iniciativas que envolvam treinamento supervisionado, frequentemente, recorrem à imagens sintetizadas para gerar pares com e sem distorção, para treinamento e avaliação dos resultados. Mesmo abordagens classificadas como não-supervisionadas ou fracamente supervisionadas, com destaque aquelas apoiadas no uso de GANs, realizam sintetização dos dados para, efetivamente, realizarem treinamento supervisionado, quer na obtenção direta dos respectivos resultados, quer em etapas constituintes do método de sua obtenção, Fabbri et al. (2018), Yu et al. (2019), Uplavikar et al. (2019), Dudhane et al. (2020). Mesmo apresentando efetividade nos resultados, estas propostas deixam a desejar quando comparadas qualitativamente com aquelas baseadas em *autoencoders* ou aquelas utilizando abordagem clássica, não baseadas em aprendizagem profunda, identificadas neste texto como abordagens por Processamento Digital da Imagem (PDI). Em adição, propostas que utilizam GAN mostram dificuldades no treinamento e na estabilidade Li et al. (2020). Isto é acentuado em modelos utilizando a arquitetura CycleGAN, onde os modelos são treinados no modo fracamente supervisionado e as imagens alvo não são explicitamente definidas (*ground-truth*) Zhu et al. (2017). Propostas recentes em DL mostram que metodologias que não necessitam de *ground-truth*, utilizando abordagens mescladas com PDI ou modelo físico, minimizam a necessidade ou não utilizam dados pareados como indicado em Li and Chen (2021); Yang et al. (2021); Saleh et al. (2022). Entretanto, ainda recorrem ao treinamento supervisionado para sub-redes que compõem seus algoritmos ou em procedimentos que realizam pré-processamento das imagens.

Metodologias baseadas no aprendizado auto-supervisionado (*self-supervised learning - SSL*) requerem baixo nível de supervisão. Define-se uma tarefa-pretexto a ser atendida pelos pro-

cedimentos de um dado método. Esta tarefa é estabelecida a partir de dados de entrada não-anotados e a sua execução requer o aprendizado de uma representação útil da imagem para ou em acordo com uma dada finalidade, Zhai et al. (2019). Um método exigindo baixo nível ou nenhum de supervisão, não exigiria dados pareados, dispensando a complexidade para a obtenção de imagens UW sintéticas sem distorção. Nas imagens sintetizadas para referência, a qualidade é dependente da técnica e do procedimento adotado para a sintetização. Ou seja, as imagens sintetizadas usadas na anotação dos dados são estimativas cuja acurácia em representar a imagem subaquática sem distorção depende do método usado para obtê-las.

Fonte de informação: Imagem de referência ou Imagem de entrada

É notório na bibliografia, o fato de que a avaliação da efetividade de um método para melhoria da qualidade de uma imagem ser bastante condicionada à percepção visual do seu resultado. Os métodos partem da imagem de entrada e a qualidade da imagem resultante é, fundamentalmente, analisada de forma subjetiva. Avaliações objetivas, baseadas em métricas, são, estimativas ou aproximações que usam como referência, imagens sintéticas ou ponderações obtidas de avaliações de *features* a partir de testes subjetivos, Li et al. (2020), Panetta and Gao (2015), Yang and Sowmya (2015). A imagem resultante de um dado método é uma estimativa de imagem melhorada, frequentemente, comparada a uma versão sintetizada, que por sua vez é, também, uma aproximação.

Neste trabalho, a imagem de entrada representa ou contém a informação real, verdadeira, disponível e a partir da qual se estabelecem as ações. Diferentemente dos métodos balizados por uma imagem melhorada e abstraindo-se a ideia de se dispor de um *ground-truth* real, uma proposta que considere utilizar como "referência" a imagem de entrada, estaria balizada pela informação real. Nestes termos, o que a imagem de entrada oferece é uma cena degradada em algum grau pelo ambiente. A degradação em si pode ser considerada como fonte de *features*, ou seja, de informação e dispondo-se de uma metodologia que a explore de forma adequada, esta poderia subsidiar sua minimização.

Conjunto de dados

A dificuldade em obter pares de imagens UW reais exige a proposição de alternativas para o treinamento dos modelos DL. A geração de *datasets* pareados a partir de versões sintéticas realçadas é amplamente explorada, Fabbri et al. (2018); Li et al. (2020); Xing et al. (2022). Entretanto, abordagens recentes apontam para o uso de estratégias com reduzido nível de supervisão (semi-supervisionado, auto-supervisão), diminuindo a dependência de dados pareados, como proposto em Li et al. (2021); Yang et al. (2021); Saleh et al. (2022). Outro aspecto bastante presente nos *datasets* empregados em métodos recentes de restauração e realce, é o volume do conjunto de dados. Percebe-se a opção dos autores por *datasets* pequenos para treinamento e, especialmente, para teste das redes neurais. Verifica-se que a quantidade de imagens destes conjuntos varia de algumas centenas até conjuntos de 10.000 imagens, Fu et al. (2021);

Lin et al. (2020); Han et al. (2021); Dudhane et al. (2020); Mello et al. (2022); Saleh et al. (2022). Embora não seja explicitamente abordado nos trabalhos, depreende-se que a redução no conjunto de dados tem conduzido à estruturas de rede mais elaboradas, com maior grau de especialização às tarefas a que se destinam, reduzindo a dependência do aprendizado em relação ao volume de dados. Em adição, destaca-se o cuidado que os autores têm demonstrado na seleção das imagens, pois as características de cena e imagem impactam na sensibilidade e desempenho dos métodos Li et al. (2020). Em função da natureza da tarefa a ser realizada, o tipo de degradação presente nas imagens UW (turbidez, *color cast*, *blurring*, redução de contraste e cor) é importante, pois contém *features* necessárias à percepção dos algoritmos e ao aprendizado das redes. Destaca-se a preocupação em manter o compromisso entre mapeamento de *features* específicas com a capacidade de generalização no aprendizado da rede.

A questão do realce de vídeo subaquático

A melhoria da qualidade de vídeo UW requer desafios semelhantes aos apresentados à melhoria de imagens, com grau de dificuldade maior, devido às exigências de consistência no tempo. A bibliografia mostra limitada disponibilidade de trabalhos em vídeo UW, sendo as iniciativas fortemente condicionadas por uma abordagem centrada no contexto "imagem", ou seja, a discussão é orientada a partir de um referencial para realce de imagem e, precariamente, estendido ao contexto de vídeo sem explorar as questões relativas à consistência temporal, Islam et al. (2020); Li et al. (2020); Saleh et al. (2022). De forma similar, trabalhos com objetivos de revisão e sintetização da discussão de metodologias para realce de imagem e vídeo UW, mostram-se bastante restritas no que tange a vídeo, Hu et al. (2022); Raveendran et al. (2021). Realce de vídeo é, consistentemente, mais desafiador que o de imagens. A consistência no tempo implica que a metodologia proposta deve garantir que o resultado da ação de melhoria da qualidade seja constante de um *frame* para outro do vídeo, evitando alterações repentinas que produzem efeito de *flickering*, bastante evidentes em sequências UW quando, por exemplo, associados à recuperação de cor e à remoção do *color cast*, Zheng et al. (2022).

Sequências de vídeo UW são coletadas de forma contínua, ou seja não apresentam cortes ou transições mostrando contextos diferentes entre *frames* sucessivos. As condições ambientais como turbidez e *color cast*, por exemplo, são constantes, e as variações são decorrentes de outras fontes como a dinâmica da cena ou do posicionamento relativo da fonte de luz. Esta constância gera limitada contribuição no mapeamento de *features* vinculadas às características da água. Isto sugere maior dificuldade na obtenção de um modelo com capacidade de generalização para vídeo e no atendimento à diversidade oferecida pelo ambiente subaquático. A proposição de estratégias de treinamento podem ser exploradas para superação destas dificuldades, como sugerido em Qi et al. (2022).

Modelo físico, PDI e realce

Na melhoria da qualidade por restauração, as implementações concorrem para obter uma estimativa da imagem UW sem degradação, balizadas pelo modelo físico de formação da imagem, Ding et al. (2019); Chang et al. (2020); Han et al. (2021). A descrição físico-matemática implícita orienta, direta ou indiretamente, a estruturação das propostas. Diferentemente, a ação de realce é executada no âmbito do espaço de cores, a partir da alteração da intensidade dos pixels. Técnicas envolvendo PDI em associação com abordagens DL são empregadas no realce de imagens Li et al. (2020, 2021) e vídeos Liu et al. (2022); Zheng et al. (2022). Este contexto analisado à luz da bibliografia relacionada ao realce, em especial nos trabalhos mais recentes, indica e permite uma flexibilidade na proposição das metodologias. Embora não seja orientada pelo modelo de formação da imagem Han et al. (2018), Li et al. (2020), a possibilidade de uma ação de realce ser inspirada no modelo físico de formação da imagem consiste em alternativa metodológica que pode ser explorada. A estimativa dos parâmetros do modelo físico é sempre etapa complexa, com várias interpretações e alternativas postas na bibliografia Wang et al. (2019); González-Sabbagh and Robles-Kelly (2023). No realce, opera-se no espaço de cores e uma alternativa inspirada no modelo físico, mas não definida por ele, sugere flexibilização neste contexto por não haver o compromisso com a estimativa dos parâmetros do modelo. Técnicas baseadas em PDI podem contribuir nesta flexibilização, auxiliando o estabelecimento de relações no espaço de cores e que sejam em algum grau, análogas ao modelo físico.

Convergência para proposta do trabalho

A questão ou problema investigado no decorrer do trabalho repousa na concepção de que a melhoria na qualidade de imagens e vídeos subaquáticos pode ser alcançada pelo condicionamento do aprendizado de uma rede neural autoencoder. Na concepção metodológica, propõe-se que este condicionamento seja performado a partir de uma penalização adicional imposta à função de perda durante a etapa de treinamento da rede. No contexto do trabalho, a penalização é representada por degradação adicional aplicada à imagem de saída da rede. Assim, espera-se que durante o treinamento a rede aprenda a decodificar em sua saída a informação codificada no espaço latente e correspondente à imagem de entrada sem esta degradação. Em adição, propõe-se que o treinamento seja realizado sem o emprego de imagens previamente melhoradas como *ground thuth*. Requer-se, portanto, que a metodologia mostre habilidade em utilizar apenas a imagem (ou *frame*) de entrada como fonte de informação, demandando, em consequência, baixo nível de supervisão. A abordagem é estruturada através da aprendizagem auto-supervisionada (SSL) tendo como propósito de ação a tarefa de realce de imagem e vídeo UW.

O arcabouço desenvolvido no trabalho usa a imagem UW original e a degradação que ela apresenta como ponto de partida. Conceitualmente, avalia-se a degradação já presente na imagem original, utilizando-a como balizamento para a geração de uma imagem sintética e

degradada da imagem de entrada. Esta versão degradada da imagem de entrada referencia o aprendizado da rede neural para efetuar a redução desta degradação e realizar o realce da imagem original.

Nos termos do aprendizado auto-supervisionado, a tarefa-pretexto corresponde, essencialmente, na remoção da degradação adicional apresentada na imagem sintetizada, executada sobre a imagem de entrada, resultando em uma imagem com realce. Importante destacar que para um resultado satisfatório, a degradação adicional deverá ser coerente com a apresentada na imagem original (entrada), ou seja, não são geradas novas *features* de degradação, mas incrementadas aquelas já presentes na imagem. Para este objetivo, a proposta apresenta uma representação para a imagem inspirada no modelo físico, mas descrita no espaço de cores a partir de características da própria imagem com o auxílio de elementos de PDI.

A abordagem do trabalho não requer dados pareados e toda a informação necessária para o realce é obtida da imagem (ou *frame* de vídeo) original de entrada. A obtenção da base de dados para treinamento do modelo é simplificada, constituindo-se em um simples conjunto com imagens subaquáticas.

O contexto vídeo requer também uma estratégia para o atendimento dos requisitos temporais de qualidade, ou seja, consistência temporal. A proposta do trabalho explora a constância das *features* ambientais como balizamento para a consistência temporal. Ao mesmo tempo, procura superar as dificuldades impostas por esta constância ao aprendizado para realce. Isto é realizado através da estratégia de treinamento em duas etapas, envolvendo imagem e vídeo.

1.2 Hipóteses Científicas

A obtenção de um modelo de degradação baseado no que a imagem já apresenta de efeitos da turbidez pode ser mais efetivo que obter um modelo de realce.

Uma rede *encoder-decoder* aprenderá a melhorar a imagem de entrada, a partir do artifício de substituir sua imagem de saída na função de perda, por uma versão com degradação aumentada e referenciada pela degradação presente na imagem de entrada.

A consistência temporal no realce de vídeos subaquáticos pode ser alcançada explorando-se a constância de *features* ambientais entre cenas de *frames* sucessivos do vídeo.

1.3 Objetivos

O presente trabalho tem como objetivo geral:

- A proposição de arcabouço para melhoria da qualidade de imagens subaquáticas através de realce e que possa ser estendido para vídeos subaquáticos, mesclando redes neurais com abordagem por aprendizado auto-supervisionado e elementos de processamento digital da imagem.

E os objetivos específicos do trabalho constituem-se em:

- Apresentar metodologia para o realce de imagens subaquáticas, fundamentada na identificação e intensificação da degradação já presente na imagem original e utilizando-a para balizar o aprendizado de um modelo DL *end-to-end*, baseado em abordagem SSL e sem a necessidade de dados anotados ou pré-processados;

- Apresentar estratégia para degradação da imagem que, simultaneamente, utilize *features* associadas à melhoria quanto à redução da qualidade da imagem, de tal forma que a intensidade desta degradação esteja em coerência com a presente na imagem original e

- Apresentar metodologia para realce de vídeo subaquático, a partir da expansão do método para imagem, propondo metodologia para o atendimento dos requisitos de consistência temporal.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são descritos conceitos importantes relacionados ao desenvolvimento do trabalho, associados ao modelo físico de formação da imagem e à arquitetura de rede *encoder-decoder*. Ao final do capítulo, é apresentada a revisão bibliográfica que fundamenta o referencial teórico para a proposta.

2.1 Modelo físico da Formação da Imagem UW

O modelo de formação da imagem subaquática descreve a degradação da cena em função da distância ao observador, estando os objetos sob iluminação atenuada a uma dada profundidade. Dois modelos, em particular, são utilizados nos trabalhos para melhoria da imagem, o proposto por McGlamery (1980) e Jaffe (1990) e o proposto por Schechner and Karpel (2005).

O modelo de MacGlamery-Jaffe estabelece que a energia luminosa E_c capturada pela câmera na água descrevendo a imagem, é uma combinação linear das componentes de transmissão direta E_d, *forward scattering* E_f e de *backscattering* E_b, indicada pela Eq. 2.1, McGlamery (1980), Jaffe (1990). Os termos da Eq.2.1 são descritos em função dos pixels, assim, o par (x, y) corresponde às coordenadas do pixel na cena.

$$E_c(x, y) = E_d(x, y) + E_f(x, y) + E_b(x, y) \tag{2.1}$$

Entretanto, o modelo depende dos parâmetros das câmeras utilizadas na aquisição das imagens, funções de espalhamento associadas ao volume de dispersão da luz, sendo utilizadas em aplicações envolvendo navegação subaquática e atividades de monitoramento, onde as informações dos dispositivos estão acessíveis Li et al. (2020).

O modelo de Schechner-Karpel, amplamente o mais utilizado, define como a cena da imagem é degradada em função da distância do objeto ao observador e condicionada pela turbidez da água e da luz ambiente à profundidade da cena. Em comparação com o modelo de MacGlamery-Jaffe, a contribuição do termo de *forward scattering* é, normalmente, desconsiderada dado que a abertura da lente da câmera é pequena, Berman et al. (2021). A Eq. 2.2 descreve o modelo de Schechner-Karpel.

$$I_{(k)}(x, y) = J_{(k)}(x, y)t_{(k)}(x, y) + {B_{(k)}}^{\infty}(x, y)(1 - t_{(k)}(x, y)), \quad k \in \{R, G, B\} \tag{2.2}$$

e

$$t_{(k)}(x, y) = e^{-\gamma_{(k)}d(x,y)}. \tag{2.3}$$

Na expressão, $I_{(k)}$ descreve a imagem subaquática capturada pela câmera, $J_{(k)}$ define a cena, $t_{(k)}$ é identificado como o fator ou mapa de transmissão e indica a razão de energia de $J_{(k)}$ capturada pela câmera e ${B_{(k)}}^{\infty}$ representa a luz ambiente. O mapa de transmissão $t_{(k)}$ é definido na Eq. 2.3, onde $\gamma_{(k)}$ corresponde ao coeficiente de atenuação da água e d a distância

do objeto à câmera.

Uma versão revisada para o modelo é apresentada por Akkaynak and Treibitz (2018). A proposta considera o efeito da dispersão direta (*forward scattering*) da luz refletida pelos objetos da cena, incluindo coeficiente de atenuação direta no termo do modelo que descreve a radiância dos objetos. Também é descrita a dependência dos coeficientes de *backscattering* e *forward scattering* com a resposta espectral da câmera que faz a aquisição da imagem. Embora o modelo apresentado permita correções na aquisições da imagens, depende de parâmetros da câmera usada na aquisição, sendo que, via de regra, não estão disponíveis.

2.2 Espaço de Cores

Áreas importantes de aplicação do processamento digital da imagem correspondem à melhoria da informação visual contida na imagem, com o objetivo de facilitar a compreensão humana, e o processamento para representação a nível de máquina, que estabelece ou permite a comunicação homem-máquina. Imagens coloridas e a percepção das cores estão presentes em todos os aspectos do cotidiano. Assim, a correta representação da cor é fundamental para viabilizar alguns destes aspectos, principalmente quando relacionado a interação homem-máquina.

Espaço de cores explicam como as cores podem ser representadas, especificando as suas respectivas componentes formadoras, definindo seus espectros de variação. Tipicamente, espaço de cores possuem entre três ou quatro componentes, sendo divididos em três categorias conforme a aplicação, Ibraheem et al. (2012): *i.* dependentes do dispositivo, quando a representação das cores são afetadas pelo *hardware* do dispositivo e devem ser compensadas (televisores, sistemas de vídeo); *ii.* Independentes do dispositivo, onde a forma de representação da cor não é afetada pelas propriedades do dispositivo, e *iii.* Orientadas ao usuário, que permitem ao usuário expressar e aproximar as cores de acordo com sua percepção. A classificação dos espaço de cores é indicadas na Tabela 1.

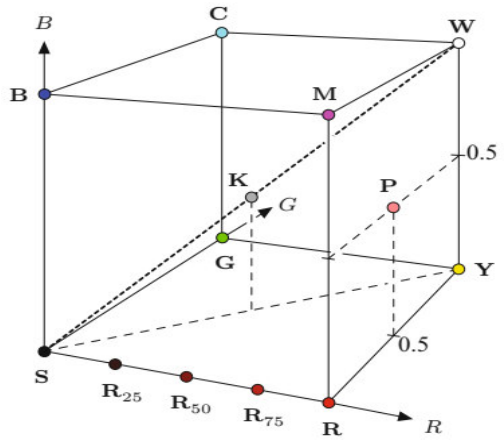
Tabela 1: Classificação dos espaços de cores. Fonte: Ibraheem et al. (2012)

Espaço de cores	**Classificação**
Munsell	Independente do dispositivo
RGB, CMY(K)	Independente do dispositivo
YIQ, YUV, YCbCr	Dependente do dispositivo
HSI, HSV, HSL	Orientado ao usuário - Dependente do dispositivo
CIE XYZ, CIE L*U*V, CIE L*A*B	Independente do dispositivo - Métrica de cor

A implementação dos modelos para melhoria de imagem e vídeo propostos foi realizada no espaço de cores RGB. O esquema de cores RGB codifica as cores como combinações das cores primárias vermelho, verde e azul. É amplamente usado em sistemas de transmissão, representação e armazenagem de imagens em cores.

O espaço de cor RGB é aditivo, ou seja, todas cores podem ser obtidas a partir da adição das cores primárias ao preto. A Fig.2.1 corresponde a representação do espaço de cores RGB através

de um cubo unitário, onde as cores primárias formam o sistema de coordenadas. Qualquer cor C_i corresponde a um ponto $C_i = (R_i, G_i, B_i)$, com o preto dado por $S = (0, 0, 0)$ e o branco por $W = (1, 1, 1)$. Os pontos localizados sobre a diagonal definida entre o preto e o branco definem escala de cinza (*shades of gray*), onde $R = G = B$.



RGB values

Pt.	Color	R	G	B
S	Black	0.00	0.00	0.00
R	Red	1.00	0.00	0.00
Y	Yellow	1.00	1.00	0.00
G	Green	0.00	1.00	0.00
C	Cyan	0.00	1.00	1.00
B	Blue	0.00	0.00	1.00
M	Magenta	1.00	0.00	1.00
W	White	1.00	1.00	1.00
K	50% Gray	0.50	0.50	0.50
$\mathbf{R}_{75}$	75% Red	0.75	0.00	0.00
$\mathbf{R}_{50}$	50% Red	0.50	0.00	0.00
$\mathbf{R}_{25}$	25% Red	0.25	0.00	0.00
P	Pink	1.00	0.50	0.50

Figura 2.1: Representação do espaço de cores RGB. Fonte: Burger and Burge (2016)

Cada canal de cor associado ao pixel é representado no intervalo $[0, 255]$. No presente trabalho, o processamento é realizado com estes valores normalizados para valores no intervalo $[0, 1.0]$.

2.3 Autoencoder

O objetivo de aprendizado não-supervisionado em DL é mapear estruturas ocultas nos dados não anotados, frequentemente, na forma de representação latente de *features* dos dados de entrada. Um autoencoder (AE) é uma rede *encoder-decoder* que realiza esta tarefa ao tentar reconstruir sua entrada. Especificamente, autoencoders são redes neurais que procuram reproduzir na saída a informação presente na entrada com o mínimo de distorção. A estrutura geral de um AE é indicada na Fig.2.2, onde n e p são as dimensões das camadas, assumindo que $0 < p < n$ para o caso geral. $\mathbf{X}$ e $\mathbf{Y}$ representam os conjuntos de vetores de entrada e saída, respectivamente.

Considerando que:

1. $\mathbb{G}$ e $\mathbb{F}$ sejam conjuntos de funções;
2. $\mathbf{A}$ seja uma classe de funções de $\mathbb{G}^p$ para $\mathbb{F}^n$ ou $A : \mathbb{G}^p \rightarrow \mathbb{F}^n$;
3. $\mathbf{B}$ seja uma classe de funções de $\mathbb{F}^n$ para $\mathbb{G}^p$ ou $B : \mathbb{F}^n \rightarrow \mathbb{G}^p$;
4. $\mathbf{X}$ seja um conjunto de m vetores de entrada (treinamento) em $\mathbb{F}$ e $\mathbf{Y}$ o conjunto correspondente de m vetores de saída em $\mathbb{F}$;
5. Δ_f é uma dissimilaridade ou distorção definida em $\mathbb{F}^n$.

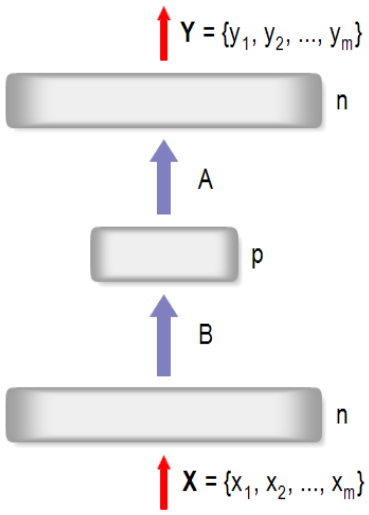


Figura 2.2: Estrutura mínima de um autoencoder

Para qualquer $A \in \mathbf{A}$ e $B \in \mathbf{B}$, a rede transforma um vetor $x \in \mathbb{F}^n$ em um vetor de saída $A \circ B(x) \in \mathbb{F}^n$. A partir do treinamento não supervisionado, o AE procura encontrar $A \in \mathbf{A}$ e $B \in \mathbf{B}$ que minimize a função de distorção global descrita pela Eq.2.4.

$$minE(A, B) = min_{(A,B)} \sum_{t=1}^{m} E(x_t) = min_{(A,B)} \sum_{t=1}^{m} \Delta_f(A \circ B(x_t), x_t) \tag{2.4}$$

O fato de ser assumido que $n > p$ corresponde à condição onde o AE tenta implementar alguma forma de compressão ou extração de *features*. Se $n \leqslant p$, onde a camada oculta é igual ou maior que de entrada, há uma solução ótima com zero distorção, envolvendo a função identidade, Baldi (2012).

Se **A** e **B** representarem os parâmetros dos segmentos de codificação e decodificação $(\theta_{enc}, \theta_{dec})$, respectivamente, e Δ_f expressar o Erro Médio Quadrático entre a entrada e a saída, e definindo-se a função de perda $\mathcal{L}$, então:

$$\mathcal{L} = minE(\theta_{enc}, \theta_{dec}), \tag{2.5}$$

$$\mathcal{L} = MSE(\mathbf{X}, \mathbf{Y}). \tag{2.6}$$

O uso de AEs se mostra bastante significativo, sendo que nas metodologias para melhoria de imagens, em especial UW, são aplicados com frequência na composição das estruturas geradoras das GANs, estruturas baseadas na U-net e em soluções próprias implementadas pelos autores.

Os autoencoders podem assumir camadas ocultas paralelas como forma de aumentar o número de representações, como mostrado na Fig.2.3. A camada expandida produz representações de tamanho $p_1 + p_2$, que podem ser apresentadas para camadas subsequentes, Baldi (2012). Diferenças entre p_1 e p_2 podem ser introduzidas por diferentes mecanismos, tais como: inicializações diferentes, algoritmos de aprendizado diferentes, diferentes amostras de treina-

mento, etc.

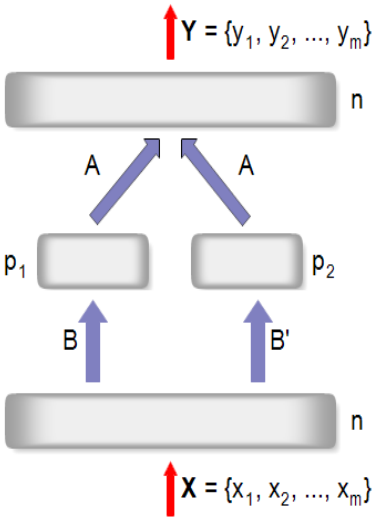


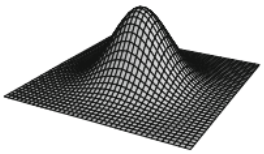
Figura 2.3: Autoencoder com camadas ocultas paralelas

2.4 Filtragem Gaussiana

O filtro gaussiano é um filtro linear de suavização da imagem, cujo *kernel* é definido a partir da função bidimensional descrita na Eq. 2.7.

$$\mathcal{G}(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{x^2+y^2}{2\sigma^2}}, \tag{2.7}$$

onde (x, y) são as coordenadas do pixel e σ corresponde ao desvio padrão da distribuição. O filtro gaussiano é isotrópico, ou seja, apresenta efeito uniforme em todas as direções. No entanto, para manter esta característica em operações discretas, recomenda-se que a matriz-*kernel* seja no mínimo de ordem 5×5, Burger and Burge (2016). A Fig. 2.4 mostra a característica do filtro e matriz-*kernel* para o caso de ordem 5×5 e desvio-padrão 1.



(a)

$$\frac{1}{273}\begin{bmatrix} 1 & 4 & 7 & 4 & 1 \\ 4 & 16 & 26 & 16 & 4 \\ 7 & 26 & 41 & 26 & 7 \\ 4 & 16 & 26 & 16 & 4 \\ 1 & 4 & 7 & 4 & 1 \end{bmatrix}$$

(b)

Figura 2.4: Filtro Gaussiano. (a) Característica 3D da distribuição do filtro gaussiano, (b) Matriz-*kernel* de ordem 5×5 e desvio-padrão 1. Fonte: Burger and Burge (2016)

O filtro gaussiano possui características de filtro passa-baixa. Assim, regiões da imagem com transições acentuadas nos valores dos pixels (alta frequência) são atenuadas. A filtragem é executada através da convolução da imagem com o *kernel*, ou seja:

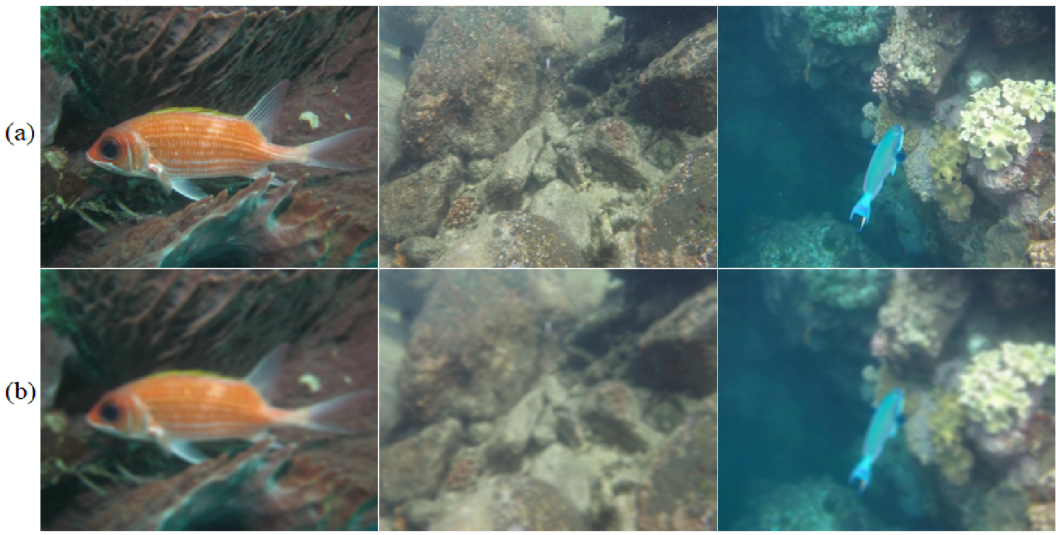


Figura 2.5: Exemplos de imagens resultantes da filtragem gaussiana. (a) Imagem original, (b) Imagem após a filtragem gaussiana

$$\mathcal{G}_H(I_s) = I_s * H_{KG}, \tag{2.8}$$

onde a I_s é uma imagem, H_{KG} é o *kernel* gaussiano e $(*)$ representa a operação convolução. Na Fig. 2.5 são mostrados exemplos de imagens resultantes da filtragem gaussiana com *kernel* 5×5 e desvio-padrão 3. Pode-se perceber a redução na percepção da textura, bordas e contornos (alta frequência), gerada pela filtragem.

2.5 Trabalhos Relacionados

Métodos para restauração e realce experimentaram grande evolução ao longo dos últimos anos. Além de novos modelos baseados em PDI, destacam-se, em especial, propostas explorando a potencialidade das redes neurais. Neste capítulo é apresentada revisão da literatura associada à melhoria de imagens e vídeos UW e que subsidiam o presente trabalho. Inicialmente, são discutidos trabalhos vinculados à restauração e ao realce de imagens. A seguir, aborda-se os trabalhos relacionados a melhoria da qualidade de vídeo. Após, abordagem sobre os *datasets* contendo imagens subaquáticas, propostos e utilizados em trabalhos recentes.

Restauração

A restauração de imagens UW usando métodos computacionais tem apresentado resultados bastante efetivos. Métodos de restauração têm seus algoritmos baseados no modelo físico de formação da imagem. A correta estimativa da luz ambiente, índice de turbidez da água e da distância do objeto à câmera é um dos objetivos principais.

O método *Dark Channel Prior* (DCP) proposto por He et al. (2010) foi originalmente destinado à imagens *outdoor* sob efeito de névoa, fumaça ou poeira (*haze*). Entretanto, foi

adaptado para cenas UW em inúmeros trabalhos. No método, a luz de fundo é calculada a partir dos 0,1% pixels mais intensos do canal mais escuro e o mapa de transmissão calculado a partir do valor mínimo da relação entre o pixel da imagem e a luz ambiente previamente estimada, definida a partir de áreas escolhidas na imagem(*patch*). Baseado neste método, Drews-Jr et al. (2013) propõe *Underwater Dark Channel Prior* (UDCP), o método considera o canal vermelho como *dark channel*, dada a maior atenuação na água, com mapa de transmissão calculado a partir dos canais verde e azul. Embora apresente resultados melhores que o DCP, os valores calculados para $t_{(k)}(x, y)$ são subestimados para objetos com valores altos nos canais azul e verde. A bibliografia traz inúmeras variantes dos métodos DCP e UDCP e são discutidas em Wang et al. (2019), Li et al. (2020), Fayaz et al. (2020). Entretanto, a ideia subjacente neste métodos não se mantém em muitas imagens subaquáticas, onde objetos próximos ao branco apresentam canais com alta intensidade nos pixels e podem resultar em baixos valores do coeficiente de transmissão, mesmo estando próximos à câmera, Berman et al. (2021).

Song et al. (2018) apresenta um modelo linear para estimar a profundidade da cena usando o conceito *Underwater Light Attenuation Prior* (ULAP). O método considera que a variação na profundidade da cena em um pixel da imagem está condicionada a diferença entre os valores máximos do canal vermelho e da intensidade dos canais azul e verde. Esta informação é utilizada para calcular o coeficiente de transmissão. O método apresenta resultados efetivos de restauração, no entanto depende de anotação prévia e manual da luz ambiente.

Em Akkaynak and Treibitz (2019) é utilizado modelo revisado proposto em Akkaynak and Treibitz (2018) para descrever o método de restauração *Sea-thru*. O método estima coeficiente de *backscattering* inspirado no DCP. A distância entre a câmera e os objetos é calculada usando a técnica do *structure-from-motion* (SFM) e o coeficiente de atenuação direta é calculado via otimização de função exponencial de dois termos descrita em função da distância. O trabalho também apresenta conjunto de imagens UW adquiridas na etapa experimental. O método considera a resposta espectral das câmeras utilizadas na aquisição das imagens. Os resultados são apresentados no contexto das imagens obtidas pelo trabalho, com reduzida turbidez e *color cast* definido.

Um *Framework* para restauração UW é descrito no trabalho de Chang et al. (2020). O fluxo do algoritmo, que é dividido em quatro etapas: correção de cor, realce de contraste local, redução de *haze* e melhoria do contraste global. A correção de cor é realizada via heurística para determinar a deficiência do canal de cor. A melhoria do contraste local é realizado aplicando-se *Contrast Limited Adaptive Histogram Equalization* (CLAHE), Zuiderveld (1994), no espaço de cores CIE-Lab. A redução de *haze* é realizada usando DCP com refinamento na determinação do mapa de transmissão através de método baseado em *image matting*. O incremento na qualidade do contraste global é executado através da expansão do histograma da imagem no espaço de cores HSI. Os resultados apresentam uma efetiva redução na turbidez, mas a habilidade na recuperação da cor não é constante.

Park and Sim (2020) descrevem algoritmo baseado na versão completa do modelo de formação da imagem, considerando o efeito de *forward scattering*. O método estima o mapa de trans-

missão a partir da distância geodésica da cor entre a cena e a luz de fundo, com refinamento usando *guided filter*. O modelo de formação da imagem é interpretado como a superposição das camadas da radiância da cena e de *backscattering*. Os parâmetros do modelo são obtidos através da minimização da função de perda relacionada a estrutura da imagem, perda de informação e o *prior* associado a luz de fundo. Os resultados mostram efetiva remoção da turbidez e melhoria do contraste das imagens. A restauração de cor apresenta média intensidade, dado que em algumas imagens há fortalecimento excessivo do vermelho. A seleção do *prior* associado a luz de fundo é feita manualmente e o processo de minimização da função de perda deve ser executado a cada imagem analisada.

Os trabalhos com propostas de restauração baseadas em DL mostram o interesse dos pesquisadores na aplicação das GANs na melhoria das imagens, valendo-se principalmente das estratégias *style-tranfer* ou *tranfer-domain*.

Yu et al. (2019) propõe método baseado em GAN para restauração UW. A rede usa Wasserstein-GAN com penalização do gradiente Zhu et al. (2017). Conjunto de imagens pareadas é obtido sinteticamente, utilizando algoritmo baseado no modelo físico e empregado no treinamento da rede. Embora seja focado em restauração, os resultados apresentados estão, qualitativamente, relacionados com realce das imagens. Trabalho similar é proposto por Wang et al. (2019), onde modelo baseado em GAN executa *style-transfer*. A GAN é treinada para gerar imagens UW sintéticas. Imagens de ambientes fechados e respectivos mapas de profundidade providas pelo *dataset NYU Depth* são utilizadas no modelo de formação da imagem para gerar versões subaquáticas sintéticas. O gerador da rede é então treinado para produzir versões UW de imagens em ambiente atmosférico. Estas são apresentadas ao discriminador que compara com imagens UW reais. O *Dataset* gerado a partir dos resultados é utilizado para treinar uma rede com arquitetura U-net, Ronneberger et al. (2015), denominada UWGAN. As imagens UW sintéticas são usadas como entrada e as imagens originais utilizadas como ground-truth.

O trabalho apresentado por Ding et al. (2019) propõe metodologia baseada em GAN semelhante à proposta anterior. Entretanto, o gerador da rede é formado por quatro módulos, responsáveis pela estimação do mapa de transmissão, luz de fundo, modelo simplificado da imagem UW e rede de preservação de borda. O *dataset* para treinamento dos módulos de estimação do mapa de transmissão e da luz de fundo é sintetizado, utilizando-se uma versão simplificada do modelo de formação de imagem, que considera a atenuação do comprimento de onda ao longo do trajeto superfície da água-objeto. São utilizadas imagens do *NYU Depth2*, pareadas com os respectivos mapas de profundidade. Durante o treinamento, versão UW destas imagens com diferentes graus de turbidez e *color cast* são obtidas empregando os parâmetros da escala de Jerlov para águas oceânicas Jerlov (1976). Os resultados demonstram a efetividade do método na recuperação da cor e contraste, mas o *dehazing* fica limitado a regiões da imagem próximas à câmera.

Arquitetura baseada em duas GANs para *dehazing* em imagens UW é apresentada por Cho et al. (2020). O trabalho utiliza *transfer-domain*, onde os geradores das GANs fazem a translação entre os domínios de imagens claras e de imagens turvas (com *hazing*). A textitloss

dos geradores é composta pela loss adversária e três termos de consistência cíclica associados à reconstrução da imagem, à filtragem laplaciana e *perceptual loss*, cujas *features* são obtidas de uma rede VGG16 pré-treinada. A função de perda dos discriminadores corresponde apenas a função de perda adversária. Para superar os problemas de baixa performance em restauração deste tipo de arquitetura, os autores propõe um algoritmo de treinamento orientado a detalhes de textura. Entretanto, os resultados mostram uma restauração de baixa intensidade das imagens.

Diferentemente dos trabalhos mencionados acima, Lin et al. (2020) propõe rede de dois estágios, dividindo a restauração UW em dois caminhos, restauração da distorção horizontal e outro da vertical. Uma rede baseada no modelo físico é responsável pela restauração horizontal, estimando o coeficiente de atenuação, luz de fundo e mapa de profundidade que são aplicados diretamente ao modelo. A distorção vertical é tratada via rede composta por função que ajusta os *features maps* dos canais de cores e por autoencoder com blocos residuais. Não são fornecidos detalhes quanto aos datasets utilizados, que são sintéticos e contém imagens representando as classes de Jerlov para águas oceânicas. A partir dos resultados mostrados, o método apresenta muito bom desempenho. Entretanto, na análise quantitativa a maioria das métricas utilizadas requerem imagem de referência, estando, assim, dependente do *ground-truth* do dataset sintetizado que não é indicado.

No trabalho de Han et al. (2021), são implementadas duas GANs para efetivarem *transfer-domain*. O modelo de formação da imagem é utilizado na determinação da imagem restaurada, embora o texto não clarifique o procedimento. O trabalho apresenta o *Heron Island Coral Reef Dataset* (HICRD), contendo doze mil imagens de oito *sites* diferentes, distribuídas em imagens de baixa qualidade, de boa qualidade e imagens restauradas. Esta últimas foram geradas pelos autores a partir das imagens de baixa qualidade, aplicando o método descrito e removendo manualmente as que não apresentaram boa restauração. O *dataset* também possui informações sobre o tipo de água, profundidade máxima da cena, atenuação em função do comprimento de onda na coluna de água e o modelo da câmera. A apresentação qualitativa dos resultados é bastante exígua e não permite uma avaliação significativa.

Desai et al. (2023) propuseram uma estrutura baseada em arquitetura *encoder-decoder* empregando *skip connections* assimétricas para restauração. O modelo usa estimativa da profundidade na cena como um *pseudo-attention* multiescala, constituindo informação adicional de entrada nas camadas do decodificar. O treinamento é realizado com pares de imagens subaquáticas sintetizadas a partir de imagens *indoor*. Na sintetização das imagens é utilizada a versão revista do modelo de formação da imagem, proposta por Akkaynak and Treibitz (2018). Os resultados mostram efetividade na recuperação da cor e remoção do *color cast*, mas com limitada redução no efeito da turbidez. Em conjunto, percebe-se nos resultados desempenho limitado na redução do borramento e recuperação da percepção de textura e bordas, mesmo com a função de perda contendo termo associado à similaridade estrutural.

Realce de imagens subaquáticas

Métodos para realce melhoram cor e contraste manipulando as intensidade dos pixels, sem, *a priori*, considerar as características das imagens UW. Abordagens iniciais aplicavam métodos de realce concebidos para imagens *outdoor*, como *Histogram Equalization*, Hummel (1977), e CLAHE, Zuiderveld (1994). Posteriormente, surgiram métodos específicos, focando características como turbidez, *color cast* e contraste reduzido, Wang et al. (2019). Estes últimos apresentam maior relevância no contexto deste trabalho, dado que utilizam os métodos anteriores em seus algoritmos e estão descritos em um ou mais espaços de cores.

Proposta fundamentada na fusão de imagens UW foi apresentada em Ancuti et al. (2012). No método, duas imagens são obtidas a partir da imagem de entrada, Uma obtida por *white balance* e outra com realce no contraste via CLAHE. A fusão entre as duas imagens é realizada por meio de um estratégia multiescala, sendo ponderada a partir do contraste e das *features* mais salientes nas duas imagens. Os resultados melhoram no contraste e a percepção de detalhes. Posteriormente, em Codruta et al. (2018) foi proposto um aprimoramento do método anterior. A imagem de entrada sofre ajuste de branco antes de ser obtidas duas imagens, uma com correção do gama e outra com acentuação de bordas. Ambas as imagens sofrem fusão multiescala, com os mapas dos pesos obtidos a partir filtragem laplaciana, estimador de saliência e do erro médio quadrático entre os canais de cores e a luminância. Os resultados mostram efetiva melhora no contraste global, acentuação de bordas e da melhor exposição de regiões escuras. Entretanto, para o balanço de branco, o método requer compensação do canal vermelho e, condicionalmente às condições da imagem de entrada, do canal azul.

Liu et al. (2016) considera o *color cast* de uma imagem subaquática produzido por uma fonte de iluminação independente e faz sua estimação utilizando o método *Deep Sparse Non-negative Matrix Factorisation* (DSNMF). O método parte da premissa de que a imagem adquirida é composta pelo produto de uma matriz de coeficientes de iluminação e uma imagem de iluminação base (iluminação canônica). A diferença entre a imagem de entrada e a versão fatorada do produto mencionado acima, é minimizada utilizando uma função de perda. O resultado fornece os coeficientes de iluminação que são utilizados na determinação da imagem com realce. O método atua sobre o *color cast* de maneira moderada, considerando o processo como resultado de um processo algébrico apenas.

Abordagem com diferentes espaço de cores é descrito por Abdul Ghani and Mat Isa (2017). Os autores propõem o algoritmo *Recursive Adaptive Histogram Modification* (RAHIM) para melhoria da cor em imagens UW. São modificadas a saturação e o brilho da imagem no espaço de cores HSV, por meio da expansão no histograma com interpolação bilinear utilizando distribuição de Rayleigh. As imagens realçadas são convertidas para o espaço de cores RGB. Os resultados são comparados com outros métodos que utilizam modificações no histograma e se mostram eficientes na recuperação de cor e contraste.

Abordagem similar utilizando expansão do histograma é desenvolvida em Huang et al. (2018), propondo o algoritmo *Relative Global Histogram Stretching* (RGHS). O método realiza

a expansão do histograma da imagem de entrada com os canais verde e azul previamente equalizados. Ruídos gerados na expansão são eliminados por filtro bilateral e a imagem filtrada é convertida para o espaço de cores CIE-Lab, onde as componentes L, a e b são expandidas em escala como forma a corrigir cor. O resultado é reconvertido para o espaço RGB. As imagens de saída indicadas no texto mostram a efetividade do método para realce.

O recente trabalho de Li et al. (2022a), está orientado para realce de cor e contraste. Na proposta são utilizados filtros bilateral e gaussiano para separação de baixas e altas frequências na imagem. A componente de baixa frequência é realçada através abordagem varacional empregando funções de peso gaussiana e *Heaviside*, utilizadas para remover os desvios na cor e realizando o realce de contraste. As imagens são convertidas para o espaço de cores HSI, dada a independência dos canais. A componente de alta frequência é filtrada via operação *soft thresholding*, Donoho (1995). Ambas as componentes são linearmente associadas para produzir a imagem realçada. Pesos para ponderação das funções gaussiana e *Heaviside*, no processamento da componente de baixa frequência, são obtidos por método de otimização com parâmetros iniciais estimados previamente. Embora as imagens de saída mostrem uma aparente tendência ao realce excessivo de contraste e do canal vermelho, os resultados destacam a efetividade do método. Os autores estenderam a análise para outros domínios além de UW, como imagens com neblina, poeira e baixa iluminação (*low light*).

Métodos para realce apoiados em DL tem recebido mais atenção dos autores que os métodos para restauração. Outro aspecto a destacar é a diversificação das arquiteturas, enquanto que para restauração as estratégias baseadas em GANs são predominantes, o mesmo não ocorre para realce, onde as GANs dividem a atenção dos pesquisadores com outras arquiteturas.

Wang et al. (2017) descreve uma rede convolucional com dois ramos, um dedicado a correção de cor e outro para remoção da turbidez. Para o treinamento da rede, um *dataset* pareado é gerado a partir da divisão em *patches* de imagens UW. Estes *patches* são degradados em diferentes níveis com a utilização do modelo físico de formação da imagem. Além das imagens pareadas, o dataset também contém os parâmetros utilizados na degradação. Resultados de saída mostram imagens realçadas, entretanto, o efeito de descontinuidade provocado pela divisão em *patches* é percebido, mesmo utilizando *guided image filter*, He et al. (2013).

Um método de correção de cor é descrito em Li et al. (2018). Conjuntos de imagens UW sintéticas são geradas a partir de uma GAN e utilizando *datasets* RGB-D formados com imagens *in air* e imagens UW coletadas em experimentos em um tanque e no mar. O gerador da GAN implementa o modelo físico de formação de imagem e utilizando a informação de profundidade da imagem RGB-D de entrada, produz uma imagem de saída sintética. As imagens sintéticas são apresentadas ao discriminador, treinado para identificar imagens subaquáticas reais. A GAN aprende a gerar versões UW das imagens *in air*. O *dataset* composto com estas imagens e os respectivos mapas de profundidade são utilizados para treinar rede *encoder-decoder* na geração de mapas de profundidade. Por fim, outro *encoder-decoder* é treinado para restaurar cor, usando como entrada os mapas de transmissão UW e imagens UW sintéticas e respectivos *ground-truth*.

Em Anwar et al. (2018) utiliza-se *dataset* sintético para treinar a rede convolucional UWCNN. Imagens UW sintéticas são geradas a partir de imagens *indoor* e empregando o modelo físico de formação da imagem, juntamente com os coeficientes de atenuação para águas oceânicas indicados por Jerlov (1976). A rede utiliza blocos residuais para melhorar a performance durante o treinamento. Nas imagens resultantes, há predominância da coloração avermelhada nos objetos da cena, indicando acentuada recuperação do canal vermelho em detrimento dos demais canais.

Em Fabbri et al. (2018) é utilizada uma GAN para realizar *style transfer* (onde duas imagens são combinadas para gerar uma terceira, que tenha o conteúdo de uma com o estilo da outra) entre imagens UW com e sem distorção. São utilizados *datasets* de imagens UW, sendo a separação entre as classes realizada manualmente com base na percepção visual. Empregando a rede Cycle-GAN, Zhu et al. (2017), são sintetizadas imagens UW distorcidas a partir de imagens sem distorção. Pares de imagens com e sem distorção são utilizados para treinar a rede UGAN, proposta pelos autores. Os resultados mostram efetiva ação de realce da rede, essencialmente, atuando na cor.

O trabalho de Li et al. (2018), propõe o uso de *transfer domain* para correção de cor. O método usa a arquitetura CycleGAN, Zhu et al. (2017), com dataset composto por imagens UW e imagens *in air*. Na função de perda, além dos termos associados à loss adversária e consistência cíclica, é utilizado termo associado a similaridade estrutural. Os autores posicionam a proposta como fracamente supervisionada (*weakly supervised*). Os resultados mostram que a rede melhora a percepção estrutural e remove *color cast* das imagens, no entanto, a recuperação de cor é deficiente.

Li et al. (2017) apresenta estratégia de fusão baseada em GAN. A rede geradora tem duas entradas uma correspondendo a imagem com distorção e outra sem distorção. As imagens são processadas por *pipelines* diferentes e a fusão ocorre na saída do gerador. O treinamento utiliza imagens pareadas, com sintetização do *ground-truth* obtido utilizando a procedimento descrito no trabalho de Fabbri et al. (2018). Os resultados mostram recuperação de cor e redução do *color cast*. O métido mostra efetividade na redução na turbidez, mas com intensidade variável em acordo com a imagem.

Metodologia baseada em rede *encoder-decoder* para restauração de cor em imagens UW é apresentada em Uplavikar et al. (2019). A representação conta com um classificador, treinado para classificar os tipos de água de Jerlov, Jerlov (1976). O treinamento é semelhante ao treinamento adversário de GANs, onde o *encoder* aprende a produzir representações latentes que conduzam o classificador ao erro. O objetivo é que *features* relacionadas a água sejam descartadas e as relacionadas à cena sejam preservadas. *Dataset* com *ground-truths* sintéticos foi produzido pelos autores seguindo o método proposto em Anwar et al. (2018).

Estratégia para redução de ruído e restauração de cor surge em Hashisho et al. (2019). A metodologia foi implementada em autoencoder baseado na arquitetura U-net, Ronneberger et al. (2015). Destaque é dado para o *dataset* sintético de imagens pareadas. Conjunto expressivo de imagens foram coletadas de diferentes fontes da internet e classificadas em imagens claras e com distorção. Via *style transfer* com uma rede CycleGAN, foram geradas imagens distorcidas

a partir das imagens claras, formando pares utilizados para o treinamento do autoencoder. A função de perda utilizada é composta por termo baseado na similaridade estrutural multiescala, Zhao et al. (2017), e outro no erro médio absoluto (norma L1). As imagens de saída mostram a efetividade do autoencoder em reconstruir imagens sem distorção. Entretanto, a distorção sintética imposta produz perda de cor, mas é, essencialmente, dada pela geração de *color cast*.

Islam et al. (2020) apresenta modelo para realce e melhoria da percepção visual, via *transfer domain* baseado em GAN com treinamento realizado em duas etapas. Um *dataset* é implementado a partir de imagens coletadas em explorações robóticas e de vídeos na internet. Estas imagens foram previamente classificadas em imagens de boa e má qualidade. Utilizando a metodologia proposta em Fabbri et al. (2018), os autores geram conjunto de imagens pareadas. Estas imagens são utilizadas no treinamento de uma rede baseada na arquitetura CycleGAN, com o objetivo de gerar imagens distorcidas a partir de imagens de boa qualidade. A rede proposta é treinada inicialmente de forma supervisionada, para gerar imagens semelhantes em *features* de alto nível ao *ground-truth*. A etapa não-supervisionada é realizada com conjunto de imagens não-pareadas, mas separadas nas imagens de boa e má qualidade. A rede aprende a mapear entre em ambos os domínios referentes à boa e má qualidade. A saída da rede é comparada à metodologias similares baseadas em GAN, apresentando melhor desempenho.

Fu et al. (2021) utiliza estratégia de normalização da imagem para reduzir a sensibilidade às *features* à água. Utilizando rede com arquitetura Unet com as *skip connections* substituídas por módulos de normalização espacial e por canal como forma de eliminar a diversidade de degradação produzida por tipos de água diferentes. Embora o treinamento seja supervisionado, os autores não informam como foi estabelecido o *ground-truth*. A função de perda é composta por dois termos, um referente ao erro médio quadrático entre as imagens e outro formado por *perception loss* cujas *features* são obtidas de uma rede com arquitetura VGG16, Simonyan and Zisserman (2015). Nos resultados mostrados, o efeito mais visível é a remoção do *color cast*.

Em recente trabalho, Li et al. (2021) propõe estratégia para realce utilizando múltiplos espaços de cores. A rede implementada tem arquitetura *encoder-decoder*. O segmento encoder é composto por três *pipelines*, referentes aos espaços de cores LAB, RGB e HSV, sendo os referentes a HSV e LAB interligados ao RGB via camadas *full-connected*. Ainda, três níveis de representação são obtidos via *downsampling*, um para cada *pipeline* e concatenados às respectivas camadas do *decoder*. Conjunto de imagens UW foi construído a partir de aquisição em diferentes locais de mergulho e com diferentes equipamentos, reunidas com imagens de *datasets* já disponíveis. O treinamento da rede é supervisionado, com *dataset* sintetizado conforme proposto em Li et al. (2020). Os autores realizaram extensa análise comparativa entre vários métodos e *datasets*. Apesar da complexidade do método, a percepção visual dos resultados mostra moderada efetividade na tarefa de realce.

Também recentemente e semelhante ao trabalho descrito acima, Yang et al. (2021) propõe rede *encoder-decoder* baseado na U-net, mas com *pipeline* contendo blocos fusão adaptativa. Estes blocos são compostos por três fluxos com número de filtros na proporção 1-3-5 nas camadas convolucionais, seguidos de concatenação e três conjuntos de camadas *full-connected*. A

função de perda é composta por termo associado a similaridade estrutural, termo associado a função de perda de Charbonnier, Barron (2019), e termo *perceptual loss*. Os autores realizaram expressiva redução no número de camadas de convolução em relação a modelo U-net original, obtendo em uma rede bastante leve. Embora na avaliação sejam mostradas imagens de entrada com reduzida distorção, os resultados mostram a efetividade do método.

Saleh et al. (2022) propuseram método para realce utilizando abordagem não-supervisionada. Um autoencoder varacional foi pré-treinado para realizar normalização adaptativa probabilística da expansão do histograma de imagens UW em dois espaços de cores. As imagens geradas pelo autoencoder são utilizadas como imagens-guias no algoritmo de supervisão para o treinamento uma rede convolucional para realce. A estratégia de treinamento emprega diversos conjuntos de dados e mostrou bom desempenho em relação aos métodos estado-da-arte. Apesar do método não requerer imagens pareadas durante o treinamento da rede proposta, o treinamento do autoencoder requer o uso imagens de referência e alto nível de supervisão.

A estrutura proposta por Cong et al. (2023) consiste de uma arquitetura GAN que utiliza o modelo físico de formação da imagem para realce UW. Uma sub-rede é usada para estimação dos parâmetros de inversão do modelo físico. O método também propõe módulo de estimação da degradação de áreas específicas da imagem, para reforço na ação de realce. O treinamento é realizado em duas etapas, ambas utilizando dados sinteticamente pareados e alto nível de supervisão. Os resultados sugerem efetividade do método, no entanto, as imagens selecionadas mostram baixo ou moderado-baixo níveis de turbidez e *color cast*. Neste sentido, os autores apontam que o método mostra limitações em realce de imagens com cenas em turbidez elevada.

Realce de vídeo

O realce de vídeos UW ainda requer maior atenção dos pesquisadores, as iniciativas disponíveis na bibliografia não definem a maturidade que o contexto de imagem apresenta. O realce de cenas subaquáticas é uma tarefa desafiadora relacionada a um problema mal-posto, Drews-Jr et al. (2016); Wang et al. (2019). As exigências de vídeo como consistência temporal e/ou desempenho em tempo real elevam a complexidade da tarefa de realce. A abordagem adotada com maior frequência é realizar realce quadro a quadro, onde cada quadro é uma imagem.

As características relacionadas a cor e luz ambiente podem variar quadro a quadro em sequências de vídeo, Hu et al. (2022). Estas variações tornam-se mais evidentes em sequências subaquáticas, dado que *frames* tendem a apresentar faixa dinâmica estreita, Mello et al. (2022), com o movimento de objetos ou da câmera pode gerar diferentes contextos de realce para quadros sucessivos, adquiridos em uma mesma condição ambiental de turbidez, profundidade e luz ambiente.

Em Li et al. (2020) foi apresentada uma rede neural bastante leve para realce de imagens UW, onde os autores propuseram aplicação em vídeo. O modelo foi treinado a partir de dados pareados, onde as imagens *ground-truths* foram obtidas sinteticamente. O conjunto de dados

foi construído usando imagens *indoor* e suas versões degradadas por meio do modelo físico de formação da imagem (Akkaynak and Treibitz (2018)) utilizando os parâmetros como proposto em Berman et al. (2021). O resultados para realce de vídeo mostraram bom desempenho no estudo comparativo. No entanto, os demais métodos considerados na comparação não foram concebidos para vídeo e, apesar dos resultados destacarem potencial para aplicações em tempo real, a análise relativa à consistência temporal do método foi explorada superficialmente.

De modo semelhante, Islam et al. (2020) propuseram modelo baseado em Rede Geradora Adversária (*Generative Adversarial Network - GAN*) também destinada ao realce de imagem, onde os autores realizaram avaliação para vídeo. O método usa abordagem de *transfer-domain* aplicada em Fabbri et al. (2018) para gerar dados sinteticamente pareados para treinamento. O modelo usa rede geradora leve e a rede discriminadora realiza o reconhecimento baseada somente em informações a nível de *patch*. Estas características atribuíram forte potencial para aplicação em vídeo, conforme os resultados apresentados. No entanto, foram relatados perda de efetividade no realce de imagens mostrando baixo contraste e perda de textura, e tendência à instabilidade durante o treinamento.

Na proposta de Zhang et al. (2023) é apresentado *dataset* elaborado a partir dos resultados de vários métodos, onde os melhores *frames* são selecionados e posteriormente refinados através da estratégia de *color transfer*. Em adição, o trabalho propõe algoritmos para realce de vídeo *online* e *offline*, utilizando mecanismo baseado em redes siamesas. Estas comparam *frame* a *frame* o vídeo alvo e o de referência e geram informação sobre qual *frame* apresenta melhor qualidade. Se o vídeo de referência apresenta melhor qualidade *color transfer* é realizada entre os quadros. A consistência temporal é obtida pela redução das diferenças de realce entre quadros durante o processamento. Embora o trabalho acuse efetividade para o método, os resultados sugerem forte atenuação das variações da média dos *frames*. Isto remete à obtenção da consistência temporal às custas de insensibilidade do realce à variações de longo prazo e que acompanham o contexto do ambiente. Infelizmente, os autores não disponibilizaram exemplos de sequências resultantes.

Paralelamente, o realce de vídeo em ambiente atmosférico (*indoor, outdoor*) mostra expressivo número de estudos, conforme relata Zheng et al. (2022). Entretanto, o realce é orientado a tópicos específicos relacionados à condições de baixa iluminação como apontado por Li et al. (2022b), *dehazing* como discutido em Peng et al. (2020) e tratamento de ruído, Huang et al. (2022). Estes trabalhos mostram foco na consistência temporal como um dos resultados da etapa de realce. Os trabalhos apresentados por Patil et al. (2020) e Zhang and Dong (2020) propõem estruturas recorrentes utilizando modelos baseados em *Long Short-Term Memory (LSTM)* para atendimento da consistência no tempo. No entanto, estas esturuturas dependem fortemente dos dados, especificamente, de *ground-truths* densos e estáveis de alto custo para obtenção, Li et al. (2021). Além disto, alguns métodos baseados em LSTM apresentaram tendência em gerar novos artefatos quando reduzindo flickering localizado, requerendo o uso de estruturas para pós-processamento para minimizar este efeitos, Thimonier et al. (2021).

Metodologias de translação vídeo-para-vídeo adaptam algoritmos para processamento da

imagem para incorporarem a dimensão tempo. O cálculo do fluxo ótico durante a etapa de treinamento dos modelos é a estratégia mais utilizada, Thimonier et al. (2021). Entretanto, estes métodos mostram arcabouços complexos, contendo diversas entradas, incluindo o quadro atual, o quadro estilizado e as respectivas saídas. Além disto, o cálculo do fluxo ótico requer alto custo computacional, com os autores procurando propondo alternativas para contornar este procedimento durante a fase de inferência dos modelos, Gao et al. (2020).

A proposta apresentada por Lei et al. (2022) explorou abordagem para o atendimento de consistência no tempo independente da tarefa de processamento vídeo propriamente dita. A metodologia identificada como *Blind video temporal consistency*, considera a consistência temporal como uma tarefa de aprendizado. O método utilizou redes LSTM convolucionais, com a função de perda contendo termo perceptual associado aos termos relativos às memórias de curto e longo prazos para,em conjunto, preservar a similaridade perceptiva associada e consistência temporal. Infelizmente, os autores não disponibilizaram sequências de vídeo mostrando o desempenho do método, embora tenham discutido várias aplicações de vídeo onde poderia ser utilizado no pós-processamento, requerendo módulos DL e estratégias de treinamento adicionais.

Datasets

Trabalhos de restauração ou realce de imagens subaquáticas baseados em DL requerem conjunto de imagens (*datasets*) para efetivar o treinamento dos modelos propostos. Estes *datasets* podem ser construídos para treinamento de modelos para aplicações bastante específicas, como avaliação de modelos ou classificação de objetos ou espécimes marinhas, por exemplo. No entanto, modelos destinados ao realce a à restauração de imagens requerem, em um contexto amplo, diversificação de cenas e *features* ambientais. Neste contexto, mesmo *datasets* destinados para uma dada finalidade, podem ser integral ou parcialmente utilizados no treinamento destes modelos, como percebido nos trabalhos de Islam et al. (2020); Lin et al. (2020); Li et al. (2020); Li and Chen (2021); Li et al. (2021); Saleh et al. (2022); Mello et al. (2022). Dada a dificuldade de se obter *ground truths* reais para imagens UW, ou seja, imagens com as mesmas cenas, mas sem a presença da água. Assim, estes conjuntos de dados não são, originalmente, não pareados. Entretanto, é frequente que propostas de realce recorram a estratégias de sintetização para obtenção de imagens melhoradas para a obtenção de conjuntos pareados, como encontrado nos trabalhos Fabbri et al. (2018); Anwar et al. (2018); Uplavikar et al. (2019); Li et al. (2021); Desai et al. (2023). A seguir são relacionados trabalhos que apresentaram os *datasets* utilizados com frequência nas propostas envolvendo realce de imagens subaquáticas.

Duarte et al. (2016) apresenta o *dataset* 3D TURBID que contém imagens UW com diferentes níveis de degradação. As imagens foram obtidas para em um ambiente controlado utilizando objetos 3D. Utilizando leite com elemento gerador da turbidez, foram adquiridas 30 imagens para 19 valores de concentração de leite, definidas entre 5 e 190 ml. Imagens de referência foram estabelecidas a partir de imagens adquiridas com a água limpa. O objetivo

é fornecer subsídios para a avaliação de métodos de restauração e para teste de algoritmos de visão computacional.

No trabalho de Fabbri et al. (2018) foram utilizados vários subconjuntos do *dataset Imagenet*, Deng et al. (2009), contendo imagens subaquáticas. A partir destes conjuntos, imagens de referências foram sintetizadas e utilizadas no desenvolvimento do modelo para realce proposto.

Em Li et al. (2020) é apresentado o *Underwater Image Enhancement Benchmark Dataset* (UIEBD), composto por 890 imagens subaquáticas reais, coletadas de modo a apresentaram diferentes características e degradação na qualidade. O *dataset* também contém imagens de referência. Os autores implementaram 12 métodos da bibliografia, cujas imagens de saída foram apresentados a 25 duplas de pessoas para a seleção com base na percepção visual. Os melhores resultados para cada imagem foram escolhidos para compor o *dataset*.

No trabalho Liu et al. (2020) é proposto o *Dataset Real-world Underwater Image Enhancement* (RUIE). Os autores distribuíram 22 câmeras em uma disposição em forma de quadrado de 10 × 10m, a uma distância de 0,5m do leito do mar, em profundidades que variaram entre 5 e 10 m. As câmeras foram ajustadas para capturar cenas entre 0,5 e 8 m, em horários pré-definidos. As imagens adquiridas foram classificadas em grupos de acordo com a finalidade de cada grupo:

- *Underwater Image Quality Set* (UIQS), com 3630 imagens e destinado a avaliação da algoritmos destinados a melhoria da visibilidade. Os autores utilizaram a métrica UCIQE, Yang and Sowmya (2015) para ranquear e classificar as imagens em cinco grupos de acordo com o *score* UCIQE;

- *Underwater Color Cast Set* (UCCS) com 300 imagens, destinado a avaliação da correção do *color cast* com imagens distribuídas em grupos com tons verde, azuis e verde-azulados;

- *Underwater Higher-level Task-driven Set* (UHTS) com 300 imagens relacionadas à vida marinha, destinado para tarefas de alto nível como classificação e identificação. as imagens foram anotadas com os tipos de vida marinha que estão indicadas nas cenas.

Islam et al. (2020) juntamente com a proposta para realce discutida anteriormente, propôs o *dataset Enhancement of Underwater Visual Perception* (EUVP) com 24.000 imagens pareadas e 6.000 imagens não pareadas. As imagens foram obtidas em explorações oceânicas, em experimentos com ROVs e de vídeos da internet. Estas imagens foram posteriormente reagrupadas, para que as imagens apresentassem variabilidade natural na sua distribuição no *dataset*. As imagens de referência para compor o segmento pareado, foram sintetizadas utilizando a metodologia proposta em Fabbri et al. (2018).

Em Islam et al. (2020), os autores propõe o *dataset* EUVP (*Segmentation of Underwater Imagery*) direcionado para tarefas de segmentação semântica UW, mas cujo conjunto de imagens subaquática pode ser aproveitado em restauração ou realce, pois contém imagens com distorção de estruturas submersas e da fauna/flora marinhas.

Considerações finais

A rápida e consistente evolução nos métodos para restauração e, principalmente, para realce de imagens subaquáticas dificulta a identificação de quais métodos representam o estado-da-arte. Esta dificuldade é bastante acentuada quando a questão relativa aos critérios de avaliação são levados em conta. A não uniformidade nas diferentes perspectivas sobre qualidade da imagem oferecidas pelas métricas e a carência de critérios para a avaliação, como por exemplo, um método ou conjunto de imagens específico para teste, são fatores que contribuem para a incerteza na avaliação de resultados.

As abordagens mais recentes mostram que as frentes metodológicas estão associadas ao processamento digital de imagens e à *deep learning*, com predomínio desta última. As tabelas 2, 3 e 4 relacionam trabalhos recentes associados à melhoria da qualidade de imagem. Em uma análise ampla, eles ensejam o estado-da-arte, considerando as respectivas metodologias e resultados. Na Tab. 5 estão relacionados trabalhos associados ao realce de vídeos subaquáticos.

Tabela 2: Métodos de Restauração (Rest.)/Realce (Rce.) com abordagens não baseadas em DL. Abreviaturas: Objetivo (Obj.), Espaço de Cores (E.C.), Processamento Digital da Imagem (PDI).

Referência	**Acron.**	**Obj.**	**Metodologia**	**Imagens**	**E. C.**	**Avaliação**
Drews-Jr et al. (2013)	UDCP	Rest.	Bas. Modelo Físico	Oceânicas	RGB	RMSE
Song et al. (2018)	ULAP	Rest.	Bas. Modelo Físico	Oceânicas	RGB	Entropia, BRISQUE, UIQM
Akkaynak and Treibitz (2019)	*Sea-thru*	Rest.	Bas. Modelo Físico revisado	Oceânicas	RGB	genérica
Chang et al. (2020)		Rest.	Bas. Modelo físico + PDI	Oceânicas, Costeiras	RGB	UCIQE, UIQM
Park and Sim (2020)		Rest.	PDI	Oceânicas	RGB	UCIQE, UIQM
Ancuti et al. (2012)		Rce.	PDI + Fusão	Oceânicas, *haze* sintético	RGB	genérica
Codruta et al. (2018)		Rce. + *dehazing*	PDI + Fusão	Oceânicas, Costeiras	RGB	PCQI, UCIQE, UIQM
Liu et al. (2016)	DSNMF	Rce.	Fatoração	Oceânicas	RGB, CIE-Lab	CRME, UCIQE
Abdul Ghani and Mat Isa (2017)	RAHIM	Rce.	PDI + interpolação	Oceânicas	RGB, HSV	Entropia, EMEE, NIQE, Contraste
Huang et al. (2018)	RGHS	Rce.	PDI	Oceânicas	RGB, CIE-Lab	Entropia, UCIQE, PSNR, MSE
Li et al. (2022a)	ACCE-D	Rce. + *Denoising*	PDI	Oceânicas, Costeiras	RGB, HSI	UCIQE, UIQM, CBPD, PCQI

Em relação às estratégias baseadas no processamento digital da imagem, destacam-se os trabalhos apresentados por Park and Sim (2020) e Li et al. (2022a), onde os resultados mostram

Tabela 3: Métodos de Restauração (Rest.)/Realce (Rce.) baseados em DL. Abreviaturas: Objetivo (Obj.), Arquitetura da Rede (Arq. Rede), Aprendizado (Ap.), Espaço de Cores (E.C.), autoencoder (AE), Supervisionado (S), Não-Supervisionado (NS), *Ground-Truth* (GT).

Referência	**Obj.**	**Arq. Rede**	**Ap.**	E.C.	***Dataset***	***Loss***	**Avaliação**
Yu et al. (2019)	Rest.	GAN, Modelo	S	RGB	Pareado, sintético	GAN	PSNR, SSIM
Wang et al. (2019)	Rest.	GAN + Unet	S	RGB	Pareado, sintético	L1, L2, SSIM, GAN	UIQM, SSIM, PSNR, MSE
Ding et al. (2019)	Rest.	GAN, CNN + Modelo	S	RGB	Pareado, sintético. Usa imagens *indoor*)	GAN, Conteúdo, gradiente, L1	SSIM, UIQM, PSNR, MSE
Cho et al. (2020)	Rest.	GANs *transfer-domain*, PDI	NS	RGB	Imagens claras e imagens UW reais	GAN, *perceptual*, Laplaciana	SSIM, PSNR, CNR, FADE
Lin et al. (2020)	Rest.	AE, CNN, Modelo	S	RGB	Imagens *indoor* pareadas. Imagens UW reais	L1, *perceptual*	UIQM, SSIM, PSNR
Han et al. (2021)	Rest.	GANs, *transfer domain*	S	RGB	Pareado, sintético c/ seleção manual	GAN	UIQM, SSIM, PSNR, MSE
Desai et al. (2023)	Rest.	AE+CNN	S	RGB	Pareado, N.pareado	MSE + MSSIM	UIQM, UCIQE, CCF, PSNR, SSIM
Wang et al. (2017)	Rce.	CNN	S	RGB	Pareado, sintético	L2	Entropia, PCQI
Li et al. (2018)	Rce.	GAN, AE, modelo	S+NS	RGB	Pareado, sintético	GAN. L2	Subjetiva
Anwar et al. (2018)	Rce.	CNN, modelo	S	RGB	Pareado, sintético	SSIM, L2	SSIM, PSNR, MSE
Fabbri et al. (2018)	Rce.	GAN, CycleGAN	S	RGB	Imagens UW sintéticas	GAN, L1, Gradiente diferencial	GDL, subjetiva
Li et al. (2018)	Rce.	GAN+CNN	S+NS	RGB	Pareado *indoor*, UW pareado sintético	GAN	RMSE
Li et al. (2020)	Rce.	GAN+Fusão	S	RGB	Pareado, sintético	GAN, L1	UCIQE, UIQM
Uplavikar et al. (2019)	Rce.	AE Unet	S	RGB	Pareado, sintético	*Loss* adversária, *Nuisance*, L2	PSNR, SSIM
Hashisho et al. (2019)	Rce.	AE	S	RGB	Pareado, sintético	L1, MS-SSIM	MSE, SSIM
Islam et al. (2020)	Rce.	GAN	S+NS	RGB	Pareado, sintético	GAN, *Feature loss*, L1	UIQM, SSIM, PSNR

Tabela 4: Métodos de Restauração (Rest.)/Realce (Rce.) baseados em DL - **Continuação**. Abreviaturas: Objetivo (Obj.), Arquitetura da Rede (Arq. Rede), Restauração (Rest.), Realce (Rce.), Aprendizado (Ap.), Espaço de Cores (E.C.), autoencoder (AE), Supervisionado (S), Não-Supervisionado (NS), *Ground-Truth* (GT).

Referência	**Obj.**	**Arq. Rede**	**Ap.**	E.C.	***Dataset***	***Loss***	**Avaliação**
Fu et al. (2021)	Rce.	AE Unet	S	RGB	Não informado	L2, *Feature Loss*	SSIM, PSNR
Li et al. (2021)	Rce.	AE, fusão	S	RGB, HSV, Lab	Pareado, sintético	*Feature Loss*, L2	SSIM, PSNR
Yang et al. (2021)	Rce.	AE	S	RGB	Pareado, sintético	Charbonnier *loss*, SSIM, *Feature Loss*	UIQM, SSIM, PSNR
Saleh et al. (2022)	Rce.	AE+CNN	NSup	RGB	Pareado, sintético - pré T.	MSE+*Feature loss*	UIQM, MUSIQ, NIQE
Mello et al. (2022)	Rce.	AE	SSL	RGB	Não Pareado	MSE	UIQM, UCIQE, NIQE, SSIM, CIE2K, PSNR
Cong et al. (2023)	Rce.	GAN	S	RGB	Pareado, sintético	GAN	UIQM, FDUM, UCIQE, CCF, MSE, PSNR

consistência na redução do efeito da turbidez, recuperação de cor e contraste. Entretanto, os métodos requerem especificação prévia e manual (empírica) de parâmetros para o processo de otimização, realizada imagem a imagem. Apesar da alta qualidade nos resultados, há forte recuperação no aspecto estrutural, atribuindo excessiva "texturização" nas imagens. Outro trabalho que pode-se destacar é o proposto por Codruta et al. (2018), que retorna resultados efetivos para realce, sendo frequentemente considerado na análise quantitativa de propostas e que apresenta boa tolerância à diversidade de imagens UW.

No contexto de restauração, destacam-se os trabalhos de Lin et al. (2020) e Desai et al. (2023). Os dois métodos se mostram bastante efetivos, mas questões associadas a restauração equilibrada da cor e redução na percepção da turbidez, muito importante para a restauração, ainda requerem atenção. Ambas as estratégias fazem uso de aprendizado supervisionado, baseada na sintetização de dados pareados, estando sujeitas à habilidade do método utilizado para a sintetização. Salienta-se o trabalho de Desai et al. (2023) faz uso de imagens em ambiente atmosférico e sinteticamente degradadas, simulando imagens subaquáticas. Estas são utilizadas no treinamento supervisionado do modelo.

No que tange a realce, há um conjunto bastante grande trabalhos que se pode relacionar como estado-da-arte, pois apresentam resultados concretos e similares. Entretanto, pode se destacar as propostas de Yang et al. (2021) e Li et al. (2021), Saleh et al. (2022) e Cong et al. (2023) que apresentam resultados expressivos. Estes métodos também utilizaram dados pareados. Entre eles, merece destaque o trabalho de Saleh et al. (2022), que mostra abordagem

não-supervisionada para realce, embora o arcabouço requeira módulo pré-treinado com dados anotados, usado na expansão adaptativa do histograma.

Tabela 5: Métodos de Realce (Rce.) de vídeos **subaquáticos**. Abreviaturas: Objetivo (Obj.), Metodologia (Metod.), Realce (Rce.), *Dehazing* (Dhz.), *Denoising* (Dns.), Aprendizado (Ap.), Consistência temporal (C.T.), autoencoder (AE), Supervisionado (S), Não-Supervisionado (NS), *Ground-Truth* (GT).

Referência	**Obj.**	**Metod.**	Metod. C.T.	***Dataset***	**Contexto**	**Avaliação**
Li et al. (2020)	Rce.	D.L. (GAN)	-	Pareado, sintético	UW	UCIQE, UIQM
Islam et al. (2020)	Rce.	D.L (GAN)	-	Pareado, sintético	UW	UIQM, SSIM, PSNR
Saleh et al. (2022)	Rce.	D.L. (AE+CNN)	-	Pareado, sintético - pré T.	UW	UIQM, MUSIQ, NIQE
Zhang et al. (2023)	Rce.	D.L. (VGG16 + Siamesa)	Pós-proc.	Imagens, videos pareados	UW	MSE, PSNR, SSIM

O realce de vídeo UW ainda carece de maior atenção dos pesquisadores. Métodos recentes dedicados ao realce de imagens subaquáticas, abordam a melhoria de vídeo como um desdobramento da avaliação do desempenho de seus algoritmos para imagem, como proposto nos trabalhos de Li et al. (2020), Islam et al. (2020) e Saleh et al. (2022). A discussão é restrita as questões de realce e considerações sobre consistência temporal não são exploradas. Diferentemente, o trabalho de Zhang et al. (2023), está direcionado para vídeo UW, apresentando metodologias para atuação *on* e *off line*, onde a consistência temporal é discutida indiretamente, no contexto da uniformidade da ação de realce e realizada como atividade de pós-processamento.

As propostas apoiadas em DL mostram que a abordagem supervisionada é a estratégia mais utilizada. Esta questão implica na disponibilidade de dados anotados, predominantemente, obtidos através da sintetização de uma imagem de referência utilizando um dado método. Isto implica na dependência de um método de realce em relação a outro. A significância do método proposto é válida, se o seu desempenho for superior ao método que é empregado para sintetizar a informação de referência. Esta consideração conduz a uma sutil contradição, quando os resultados obtidos utilizando-se os dados de referência sintetizados, não são, explicitamente, comparados com aqueles resultantes do método utilizado na sintetização desta referência. Isto ocorre na maioria dos trabalhos.

Outro ponto que se observa está relacionado a complexidade estrutural dos modelos. Embora seja de difícil quantização e não considerado nas avaliações, o aporte computacional necessário, a estabilidade, tempo de treinamento e volume de dados são exemplos que permitem a percepção subjetiva da complexidade de um modelo.

A degradação original imposta pelas condições da água e iluminação às cenas de imagens e vídeos constitui-se em fonte de informação, não observada ou considerada de forma objetiva na ampla maioria dos trabalhos relativos ao realce, e tratada de forma indireta na restauração,

a partir dos mapas de profundidade e estimativas da turbidez da água.

Nesta linha, o problema mal-posto definido pela melhoria da qualidade da imagem UW, permite que sejam exploradas alternativas na proposição de soluções. A imagem ou vídeo original ou de entrada, consiste na informação real, efetiva, concreta do que se observa. Identificar a natureza da degradação presente na entrada poderia ser explorada como informação de referência, e a consequente estratégia para melhoria da imagem.

A partir da bibliografia, verifica-se que a investigação para se alcançar consistência temporal, envolvendo vídeos no ambiente atmosférico, são mais frequentes e mostram maior evolução em relação ao ambiente subaquático. Neste contexto, a consistência temporal eleva o desafio metodológico, pois consiste em tarefa adicional ao realce. Proposta neste linha de investigação são importantes para o desenvolvimento da área.

3 ASPECTO CONCEITUAL DA PROPOSTA

Este capítulo é dedicado à descrição da proposta para melhoria de imagens subaquáticas a nível conceitual. A metodologia proposta considera duas premissas associadas à imagem UW e que permeiam todo o trabalho. A primeira premissa assume que toda a imagem subaquática apresenta degradação em algum contexto ou nível. Por degradação, entende-se a presença de elementos na cena da imagem que limitem ou prejudiquem a percepção visual de cor, contraste, bordas, contornos e textura dos objetos e/ou causem distribuição irregular da luz incidente na cena. As condições ambientais em que a cena se encontra definem, na maioria dos casos, a degradação. É condicionada, principalmente, à água, ao tipo e concentração do material em suspensão, à profundidade da cena, à intensidade e posição relativa da fonte luz. A segunda premissa considera a degradação como fonte de informação, onde o conhecimento e estimativa desta informação podem contribuir para remoção ou redução da degradação percebida na cena, auxiliando na tarefa de melhoria da qualidade.

A relação entre ambas as premissas subsidia a mudança na perspectiva metodológica proposta para a efetivação da ação de realce sobre imagens subaquáticas. Em essência, um modelo baseado em aprendizado profundo requer que sua informação de saída, obtida a partir de uma informação de entrada, seja comparada a uma referência durante o treinamento. Ou seja, os dados utilizados no treinamento da rede (modelo) devem ser constituídos em pares, representando a entrada e a referência. No contexto da melhoria de imagens, especificamente, no realce de imagens UW, é comumente adotado a representação destes pares pela imagem subaquática de entrada, mostrando a cena com degradação definida em função das condições ambientais, e uma versão desta, com qualidade melhorada (realçada), que configura a informação de referência.

No entanto, versões melhoradas, empregadas como referência, exigem método prévio para obtê-las e a respectiva qualidade depende deste método. Até que ponto a cor percebida de um objeto em uma cena subaquática corresponde à correta? Qual a influência do *color cast*, profundidade e luz ambiente? Importante destacar que a informação real, verdadeira de cor, contraste e iluminação em uma imagem subaquática não está disponível na quase totalidade dos casos. O uso de imagens pareadas no arranjo entrada e versão melhorada, tensiona qualquer nova proposta para realce. Pois, a reflexão sobre a justificativa para uma nova proposta para realce UW empregando redes neurais, aponta que seus resultados devem ser melhores que as versões realçadas utilizadas no pareamento dos dados. Ou seja, seus resultados devem ser melhores que o proposto como referência para o treinamento. Esta linha de análise sugere um potencial paradoxo na justificativa para um método de realce UW, quando utilizando imagens realçadas através de um método prévio. Esta condição demanda aumento de complexidade e impõe dificuldades à avaliação dos resultados. Percebe-se que esta questão tem sido sutilmente evitada na bibliografia, através da valorização dos métodos em detrimento dos resultados.

A metodologia proposta neste trabalho assume abordagem diferente quanto aos dados utilizados no treinamento do modelo, onde a imagem de entrada é empregada como referência, ao invés do uso de imagens previamente realçadas. O objetivo é favorecer o aprendizado do

modelo (rede) na identificação e na consequente minimização das características da degradação da imagem. A estratégia é baseada na intensificação da degradação percebida na imagem, que atua como viés condicionador do aprendizado para a minimização desta degradação pela rede durante o treinamento.

3.1 Conceito

O método consiste em conduzir o aprendizado de uma rede neural (autoencoder) pela modificação da informação que é, efetivamente, apresentada à função de perda durante o treinamento. A Fig. 3.1 ilustra o método no seu aspecto conceitual.

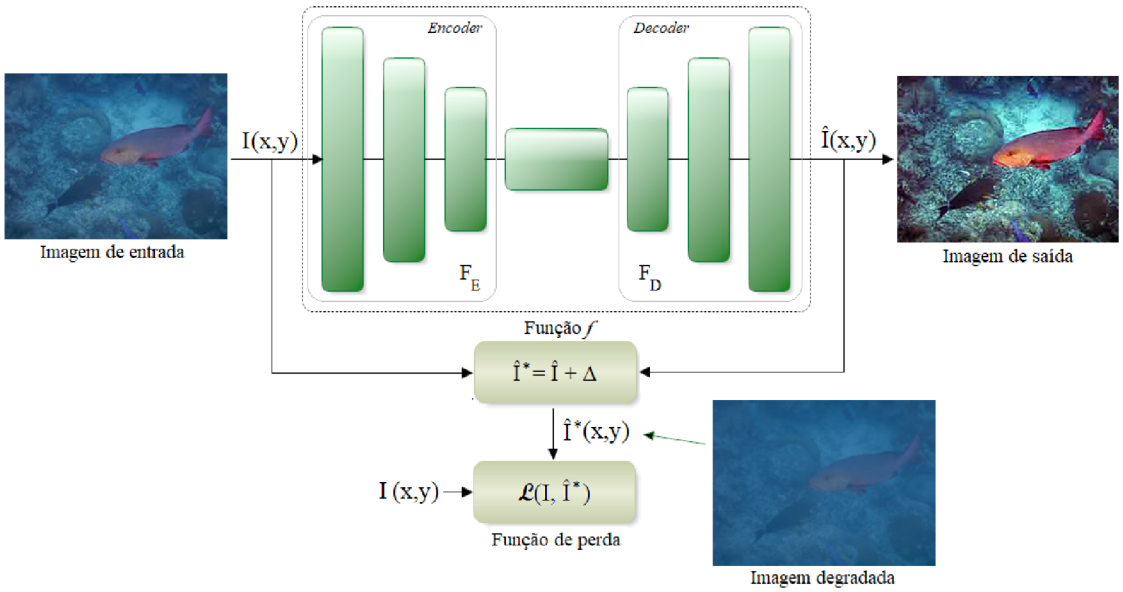


Figura 3.1: Aspecto conceitual do método

Uma imagem UW de entrada I de um autoencoder é decodificada durante o treinamento, apresentando na saída a imagem $\hat{I}$. Antes de ser apresentada à função de perda, $\hat{I}$ é modificada pela função f. As *features* relacionadas à degradação já presente na cena de I são intensificadas por f, gerando a versão $\hat{I}^*$ degradada de I. Ou seja:

$$\hat{I}^* = f(\hat{I}, I). \tag{3.1}$$

Conceitualmente, uma degradação adicional Δ representando as características (*features*) de degradação presentes em I é imposta a $\hat{I}$, resultando na imagem $\hat{I}^*$ com degradação intensificada. Ou seja:

$$\hat{I}^* = \hat{I} + \Delta. \tag{3.2}$$

No método, $\hat{I}^*$ substitui $\hat{I}$ na função de perda (L), sendo assim comparada à I. A penalização imposta pela função de perda, orienta o aprendizado da rede a compensar a degradação adicional

Δ definida em $f(\hat{I})$, durante o treinamento. Como resultado, a rede decodifica $\hat{I}$ como uma versão de I sem a degradação Δ. Nas expressões a seguir, o par (x, y) representa as coordenada do pixel na imagem.

A imagem de entrada $I(x, y)$ é codificada pela seção *encoder* F_E, gerando uma representação latente da entrada. O *decoder* F_D decodifica a informação latente na versão similar $\hat{I}(x, y)$ da entrada. Ou seja,

$$\hat{I}(x, y) = F_D(F_E(I(x, y)). \tag{3.3}$$

Ao longo do treinamento $\hat{I}$ é alterada pela função $f(I)$, produzindo a imagem degradada $\hat{I}^*(x, y)$, conforme Eq. 3.4, que é comparada a I na *loss*.

$$\hat{I}^*(x, y) = f(\hat{I}(x, y)). \tag{3.4}$$

Considerando θ_E e θ_D representem os parâmetros da rede, atualizados durante o treinamento de acordo com a função de perda L, ou seja:

$$\theta_E, \theta_D = f(min[L(I(x, y), \hat{I}^*(x, y))]). \tag{3.5}$$

Se a degradação imposta por $f(\cdot)$ for similar à presente em I, então minimizando a diferença entre I e $\hat{I}^*$, espera-se que a rede produza uma versão restaurada I' da entrada. Para verificar-se esta expectativa, pode-se considerar como uma questão de problema inverso ligado à imagens, Tamir et al. (2020). Considerando o modelo linear com distorção aditiva:

$$\mathbf{y} = \mathcal{K}\mathbf{x} + \epsilon, \tag{3.6}$$

onde $\mathbf{y} \in \mathbf{R}^n$ é a informação (imagem) com ruído, $\mathbf{x} \in \mathbf{R}^m$ é a informação a ser medida (imagem adquirida), ϵ é o ruído presente na medida e $\mathcal{K} \in \mathbf{R}^{m \times n}$ representa o sistema de medida. A metodologia do problema inverso estabelece que pode-se obter uma estimativa $\hat{\mathbf{x}}$ da informação $\mathbf{x}$, desde que $\hat{\mathbf{x}}$ contenha um mínimo $h < m$ de valores diferentes de zero (esparsidade limitada), Ongie et al. (2020), Amjad et al. (2018). Esta estimativa pode ser obtida através da minimização por mínimos quadrados da expressão:

$$\hat{\mathbf{x}} = min_x \frac{1}{2} ||\mathbf{y} - \mathcal{K}\mathbf{x}||_2^2. \tag{3.7}$$

Assim, no contexto do trabalho, podemos reescrever a Eq.3.6 como:

$$\hat{I}^* = \mathcal{K}I + \Delta(I) + \epsilon, \tag{3.8}$$

onde $\Delta(I)$ corresponde à degradação presente na imagem de entrada. $\mathcal{K}$ representa a reconstrução da imagem de entrada pelo autoencoder, assim $\hat{I} = \mathcal{K}I$. Incorporando o ruído ϵ à degradação Δ, tem-se:

$$\hat{I}^* = \hat{I} + \Delta(I), \tag{3.9}$$

Como a dimensionalidade das imagens de entrada e saída são idênticas e $\mathcal{K}$ pode ser aproximado pela função identidade (autoencoder), não há problema devido à esparsidade e o sistema apresenta uma única solução. Se a função de perda de reconstrução for implementada via função de Erro Quadrático Médio (*Mean Square Error* - MSE), tem-se:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} ||I_i - \hat{I}_i^*||^2, \tag{3.10}$$

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} ||I_i - (\hat{I}_i + \Delta_i(I))||^2, \tag{3.11}$$

A imagem de entrada pode ser representada por $I = I' + \Delta(I)$, onde I' é a imagem sem degradação, substituindo na Eq.3.11:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} ||I_i' + \Delta_i(I) - (\hat{I}_i + \Delta_i(I))||^2. \tag{3.12}$$

Ao longo do treinamento a função de perda da Eq.3.12 é minimizada, ou seja:

$$min \sum_{i=1}^{N} ||I_i' - (\hat{I}_i)||^2. \tag{3.13}$$

Assim, com a minimização tem-se que:

$$\hat{I}_i \rightarrow I_i', \tag{3.14}$$

Ou reescrevendo a Eq.3.13 com $\hat{I} = \hat{I}' + \hat{\Delta}(I)$:

$$min \sum_{i=1}^{N} ||I_i' - (\hat{I}_i' + \hat{\Delta}(I))||^2. \tag{3.15}$$

A minimização impõe que:

$$\hat{I}_i' \rightarrow I_i' \quad e \quad \hat{\Delta}(I) \rightarrow 0. \tag{3.16}$$

As expressões 3.15 e 3.16 mostram que a imagem reconstruída pelo autoencoder ao final do treinamento e no contexto da metodologia proposta, tende a apresentar menor degradação que a entrada (idealmente seria nula). A expressão 3.16 salienta que quanto melhor for a descrição ou representação de Δ, maior será a redução na degradação sobre a imagem de saída $\hat{I}$.

A metodologia não requer *dataset* com imagens pareadas, permitindo níveis menores de supervisão como em SSL. Nestas abordagens, o objetivo é o mapeamento de estruturas ocultas em dados não anotados, frequentemente, na forma de representação latente de *features* dos

dados de entrada, Baldi (2012). O autoencoder realiza esta tarefa, procurando reproduzir na saída a informação presente na entrada com o mínimo de distorção.

A imagem de entrada assume a condição de imagem de referência durante o treinamento da rede. Assim, há a alteração na perspectiva do *estima-se para onde ir* posta pelo uso de imagens previamente realçadas como referência em dados pareados, para *sabe-se de onde partir*, contexto definido pela utilização da imagem de entrada como referência.

Importante destacar que a efetividade do método está relacionada à degradação imposta em $f(\hat{I})$, cuja estimativa deve ser coerente com a degradação já presente na cena de I. A expressão 3.16 salienta esta condição, pois quanto melhor for a descrição ou representação de Δ, maior será a redução na degradação sobre a imagem de saída $\hat{I}$.

3.2 Função de degradação

A redução na qualidade da imagem a ser executada pela função f, Figuras 3.1 e **??**, deve ser coerente com aquela apresentada na cena da imagem de entrada. Recuperando a Eq. 3.1 na Eq. 3.17, tem-se:

$$\hat{I}^*(x,y) = \hat{I}(x,y) + \Delta(x,y). \tag{3.17}$$

As imagens $\hat{I}^*(x,y)$ e $\hat{I}(x,y)$ em 3.17 são descritas em termos dos pixels de coordenadas (x,y). Em consequência, o incremento Δ representando a degradação, também é descrito em termos de pixels. Em outras palavras, Δ é uma imagem, contendo a informação relativa à degradação presente na imagem de entrada I e descrita a nível de pixel.

A geração da imagem incremental de degradação Δ é realizada a partir da interação entre duas versões da imagem de entrada. Conceitualmente, uma versão sem distorção I_e e uma versão com distorção acentuada I_d. Assume-se que ambas as versões diferem de I por uma imagem incremental Δ_i, ou seja:

$$I_e(x,y) = I(x,y) - \Delta_i(x,y), \tag{3.18}$$

e

$$I_d(x,y) = I(x,y) + \Delta_i(x,y). \tag{3.19}$$

A partir da diferença entre I_d e I_e, obtém-se:

$$I_d(x,y) - I_e(x,y) = \Delta(x,y), \tag{3.20}$$

onde:

$$\Delta(x,y) = 2\Delta_i(x,y). \tag{3.21}$$

Substituindo Δ na Eq. 3.17 pela expressão indicada na Eq. 3.20, obtém-se:

$$\hat{I}^*(x,y) = \hat{I}(x,y) + I_d(x,y) - I_e(x,y). \tag{3.22}$$

As descrições ao longo de trabalho são realizadas em termos dos canais de cores vermelho, verde e azul (RGB). Assim, pode-se reescrever a Eq. 3.22 generalizando para os canais de cores:

$$\hat{I}^{*}_{(k)}(x,y) = \hat{I}_{(k)}(x,y) + I_{d(k)}(x,y) - I_{e(k)}(x,y), \tag{3.23}$$

onde k $\in \{R, G, B\}$.

A Eq. 3.23 sumariza a função de degradação adotada neste trabalho, onde $\hat{I}^{*}_{(k)}$ é a imagem resultante da função f indicadas nas Figs. 3.1 e **??**.

À primeira vista, a interpretação para a Eq. 3.23 aponta que a degradação é obtida da diferença entre as imagens I_d, que contém *features* associadas à redução da qualidade, e a I_e que contém *features* associadas à melhoria a imagem. No entanto, propõe-se uma interpretação alternativa que subsidie o entendimento do propósito de ambas as imagens, envolvendo o contexto do realce como problema mal-posto. Nesta concepção, as *features* que favorecem incertezas na percepção de cor, contraste e detalhes de textura tenham degradação fortalecida em I_d. Por outro lado, elementos que não gerem incertezas na sua percepção (*features* que não contribuem com a degradação), como exemplo os elementos estruturais e até mesmo situações envolvendo cor, são valorizados em I_e. Em termos abrangentes, ambas as perspectivas para interpretação da degradação, expressas a partir de I_d e I_e, exploram a aparente dualidade entre estas imagens, onde o fluxo desenvolvido para a geração de uma priorize *features* que tenha contexto complementar à outra.

4 METODOLOGIA PARA MELHORIA DA QUALIDADE DE IMAGENS SUBAQUÁTICAS

A melhoria da qualidade de imagens subaquáticas concebida a partir de uma abordagem por aprendizado profundo, encontra na falta de dados pareados um dos seus maiores desafios. Esta condição é significativa para métodos baseados em aprendizado profundo, pois depende, fundamentalmente, da base de dados para a extração de informações para a efetivação da melhoria. A disponibilização de dados pareados é interessante, pois admite nível maior de supervisão no treinamento dos modelos. Em um contexto amplo, isto facilita a concepção da estratégia a ser implementada, com as imagens de referência fornecendo informações adicionais relativas à cor, contraste e intensidade da luz ambiente, por exemplo. Entretanto, devido a dificuldade e ao custo para se obter um *dataset* contendo *ground-truths* de imagens subaquáticas, várias propostas optaram pela sintetização de imagens com melhor qualidade visual para uso em seus respectivos métodos.

A sintetização implica em ter a disposição ou implementar método para a melhoria de imagens subaquáticas, que serão utilizadas como imagens de referência na construção de uma base de dados (imagens) anotada. Os dados anotados resultantes estão, assim, disponíveis para serem utilizados no desenvolvimento de novos métodos. Coerentemente, o contexto formado sugere que resultados (imagens) de um novo método, que utilize estes dados pareados no treinamento dos modelos, deveriam ser melhores aos utilizados como referência. Em adição, estes resultados são, essencialmente, dependentes dos previamente sintetizados.

A percepção real da cor (tons e intensidades) e contraste, por exemplo, não são conhecidos na maioria dos casos envolvendo a melhoria de imagens subaquáticas. Imagens de referência sintetizadas, e os resultados obtidos a partir delas, oferecem uma versão das condições reais da cena, cuja qualidade depende dos métodos utilizados.

O modelo proposto no presente trabalho explora a degradação presente na cena da imagem UW como fator de contribuição na sua remoção na tarefa de realce. Isto implica que o modelo seja efetivo no tratamento da turbidez, *color cast* e nos efeitos distribuição irregular da luz ambiente. Assim, a informação sobre degradação, ou seja, o conjunto de *features* que representam a degradação pode ser segmentado quanto à contribuição ao enfoque do realce como problema mal-posto em dois contextos. Um contexto agrupando *features* que efetivamente contribuem com a "incerteza" característica do problema mal-posto e que impacta na percepção de cor e contraste, especificamente, turbidez, *color cast* e as respectivas interações com a luz ambiente. Outro contexto agrupa elementos da degradação que não geram "incertezas" significativas à percepção da cena, como o borramento (*blurring*) sobre elementos estruturais como bordas, contornos e elementos de textura. A percepção destes elementos de estrutura tendem a ser agnósticos quanto à turbidez e quanto ao *color cast*, desde que não provoquem oclusão total de detalhes ou objetos na cena. Estes contextos são concomitantes e contemplados nas imagens I_d e I_e da Eq. 3.23. Entretanto, o conjunto de *features* de degradação está presente em

ambas as imagens I_d e I_e, sendo, no entanto, que os contextos destas *features* são valorizados diferentemente em uma imagem em relação a outra.

Importante o fato da proposta não requerer o uso de imagens previamente realçadas na composição da base de dados. Por premissa metodológica, a imagem UW consiste na única fonte de informação disponível e assume-se que ela contenha informações suficientes para viabilizar a ação de método.

Nas seções a seguir é apresentada a metodologia para melhoria da qualidade desenvolvida no trabalho e viabilizada através de uma estratégia para realce de imagens subaquáticas.

4.1 Módulo de Degradação

A função f nas Figuras 3.1 e **??** representa o aspecto conceitual para a obtenção da imagem degradada $\hat{I}^*$. Entretanto, o algoritmo proposto para a degradação compreende um conjunto de transformações sobre a imagem de entrada, compondo um todo, cujo resultado final é a operação descrita pela Eq. 3.23. Este conjunto de operações que concorrem para a geração de $\hat{I}^*$ será identificado ao longo do restante do texto, como Módulo de Degradação (MD).

A Fig. 4.1 ilustra a estrutura do MD. O módulo recebe como informação de entrada a imagem original $I_(k)$. Dois fluxos (*pipelines*) duais processam a imagem original com o objetivo da geração de duas imagens-guias. Um *pipeline* é responsável pela geração da imagem guia para realce, identificada por I_e, e o outro responsável pela geração da imagem-guia para degradação I_d. Objetivamente, I_e congrega *features* "boas", associadas à percepção da qualidade da imagem, enquanto I_d inclui *features* "ruins", relacionadas com degradação e perda de qualidade. Salienta-se que ambos os aspectos (duais) contribuem para a formação da imagem degradada $\hat{I}^*$, que é obtida pelo acréscimo de *features* "ruins" e pela redução nas *features* "boas". A seguir são detalhados os blocos constituintes do MD, funcionalidades e as respectivas imagens intermediárias geradas no processo.

4.2 Mudanças no histograma

A partir da imagem de entrada, cada *pipeline* do MD gera uma versão inicial da respectiva imagem-guia. Estas versões iniciais são determinadas a partir da expansão do histograma (*histogram stretching*) e sua operação inversa expressa como contração do histograma (*histogram shrinking*).

A expansão do histograma é um procedimento regularmente utilizado em trabalhos associados à melhoria de imagens, especialmente, em imagens submersas, como percebido nos trabalhos de Abdul Ghani and Mat Isa (2017); Anwar et al. (2018); Chang et al. (2020); Li et al. (2022a); Saleh et al. (2022). O procedimento melhora o contraste global e o realce de cores, permitindo uma percepção mais natural da cena, Abdul Ghani and Mat Isa (2017). A expansão do histograma produz efeitos mais perceptíveis sobre imagens UW do que quando aplicada a imagens com cenas em ambiente atmosférico. A expansão produz cenas mais agradáveis à observação

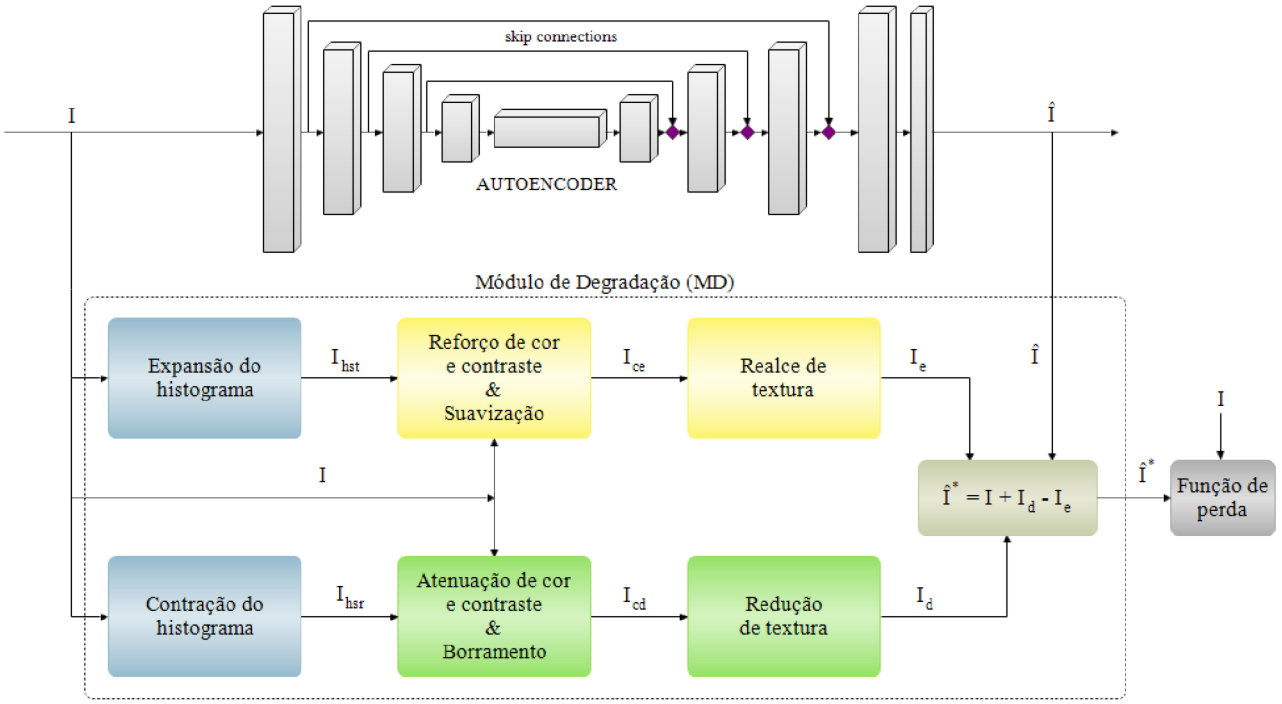


Figura 4.1: Diagrama mostrando a estrutura do Módulo de Degradação

visual. Isto é devido a tendência das imagens UW apresentarem faixa dinâmica reduzida. A contração do histograma é obtida pela operação inversa à expansão e resulta em uma imagem com redução perceptível no contraste e cor.

As funções de expansão e contração do histograma são expressas. respectivamente, nas Eqs. 4.1 e 4.2 :

$$I_{hst(k)}(x,y) = \frac{I_{(k)}(x,y) - \min(I_{(k)}(x,y))}{b_{w(k)}}, \tag{4.1}$$

e

$$I_{hsr(k)}(x,y) = I_{(k)}(x,y)b_{w(k)} + \min(I_{(k)}(x,y)), \tag{4.2}$$

onde $I_{hst(k)}$ e $I_{hsr(k)}$ são as versões da imagem de entrada $I_{(k)}$ com histograma estendido e contraído, respectivamente. $\min(I_{(k)})$ corresponde ao valor mínimo por canal de cor da imagem e $b_{w(k)}$ corresponde a faixa dinâmica da imagem, expressa por:

$$b_{w(k)} = \max(I_{(k)}) - \min(I_{(k)}), \tag{4.3}$$

onde $\max(I_{(k)})$ representa os valores máximos por canal de cor. Os resultados para ambas as operações são mostrados na Fig. 4.2, para um exemplo de imagem UW e as versões obtidas pelas Eqs. 4.1 e 4.2.

Além das diferenças visuais no aspecto de cada imagem, os respectivos histogramas permitem avaliar o efeito da redução e do aumento na percepção da degradação. O histograma da versão com expansão, Fig. 4.2(b), mostra os canais de cores com intensidades distribuídas com relativa uniformidade. No sentido oposto, o histograma da Fig.4.2(c) mostra o incremento separação dos canais, com o vermelho restrito a baixas intensidades, o azul em tonalidades

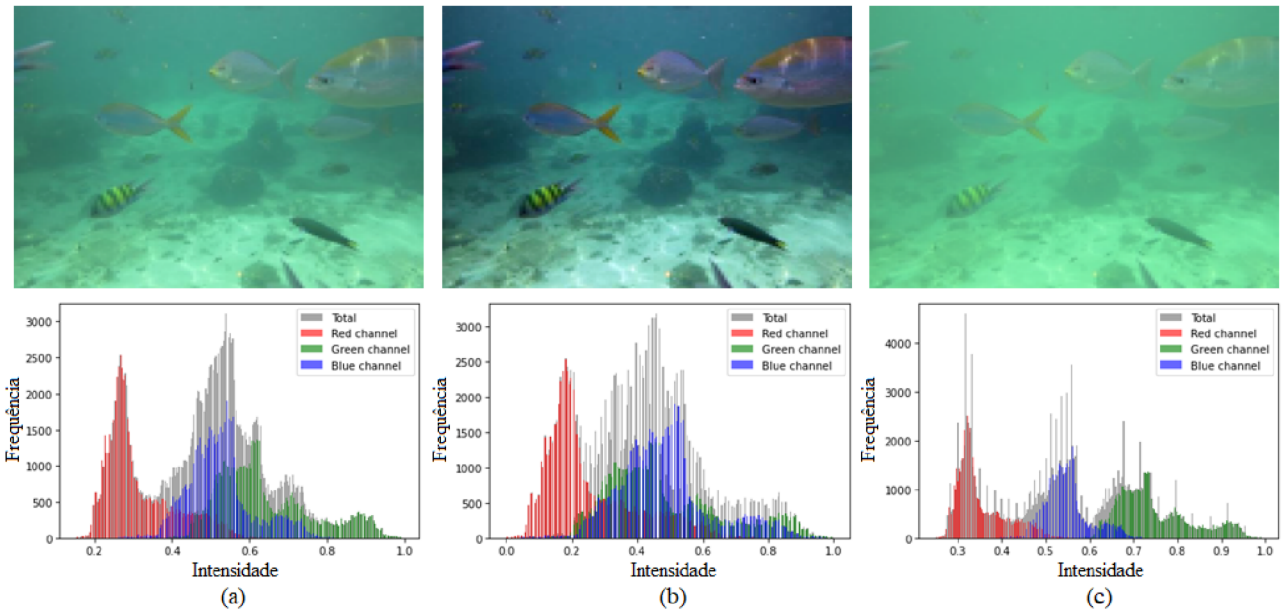


Figura 4.2: Imagem UW com suas versões com histograma expandido e contraído e respectivos histogramas. (a) Imagem original $I_{(k)}$, (b) Versão com expansão do histograma $I_{hst(k)}$ e (c) Versão com contração do histograma $I_{hsr(k)}$.

intermediárias e o verde nas mais intensas, acompanhando a percepção visual do aumento do *color cast*.

4.3 Ajustando cor e contraste

A expansão do histograma consiste no deslocamento da distribuição dos pixels no histograma, seguido de normalização na faixa dinâmica. O efeito como mostrado na Fig. 4.2(b) é uma melhor percepção visual da cena. Entretanto, a expansão apenas incrementa a informação de cor disponível na imagem. Em adição, devido a normalização, há a produção de intervalos (*gaps*) sem informação alguma de cor (tons) no histograma. Assim, se a ação de realce for realizada com base apenas na imagem com o histograma expandido, há limitação no resgate de cor (e contraste), principalmente se a imagem original apresentar estreita faixa dinâmica. Efeito complementar ocorre na contração do histograma, onde há redução da faixa dinâmica seguida de deslocamento por incremento dos valores do pixels no espaço de cores. Como resultado, há a redução na largura de faixa da imagem resultante e desequilíbrio na distribuição dos pixels ao longo dos canais de cores, como se espera obter no *pipeline* de degradação. No entanto, há perda de informação de cor (e contraste), devido, principalmente, ao desequilíbrio produzido nos canais de cores. Este efeito é acentuado quando a imagem original apresenta estreita faixa dinâmica.

A estratégia para minimizar os efeitos descritos emprega uma representação da imagem no espaço de cores, inspirada no modelo físico de formação da imagem proposto por Schechner and Karpel (2005) e discutido por Akkaynak and Treibitz (2018), amplamente utilizado nos trabalhos relacionados à restauração. A expressão para o modelo é indicada na Eq. 2.2 e, por

conveniência, é reescrita na Eq. 4.4 com $t_{(k)}(x, y)$ substituído conforme a Eq. 2.3.

$$I_{i(k)}(x, y) = J_{(k)}(x, y).e^{-\gamma_{(k)}d(x,y)} + (1 - e^{-\gamma_{(k)}d(x,y)}).B_{(k)}, \tag{4.4}$$

onde $I_{i(k)}$ é a imagem subaquática descrita pelo modelo, $J_{(k)}$ a informação (luz) refletida pelos objetos na cena, $\gamma_{(k)}$ é o índice de atenuação da luz (turbidez), d é distância dos objetos à câmera (sensor) e $B_{(k)}$ representa a luz disseminada no ambiente, gerada pela colisão dos raios de luz da fonte com as partículas do material em suspensão na água, ou *veiling light*, de acordo com Schechner and Karpel (2005). Essencialmente, o modelo descreve a imagem como formada pela interação no sensor da luz refletida dos objetos e a luz ambiente. Ou seja, a imagem é descrita por um termo associado à radiância dos objetos sob o efeito da turbidez, denominado de *forward scattering* e descrito pelo termo $J_{(k)}(x, y).e^{-\gamma_{(k)}d(x,y)}$. O segundo termo, $(1 - e^{-\alpha_{(k)}d(x,y)}).B_{(k)}$, se relaciona aos raios de luz da fonte desviados em direção ao sensor, a partir de colisões com as partículas do material em suspensão na água, denominado de efeito *backscattering*, Akkaynak and Treibitz (2018). As informações de luz propostas por ambos os termos concorrem no sensor, como sugerido pela complementaridade dos fatores envolvendo as exponenciais na Eq. 4.4.

A representação da imagem proposta é matematicamente similar ao modelo físico (Eq. 2.2), mas descrita no espaço de cores, ou seja, todos os seus parâmetros são descritos a partir dos valores dos pixels da imagem representada, como indicado na Eq. 4.5.

No método, todas as expressões envolvendo imagens são descritas em termos de pixels, cujas coordenadas são (x, y). Entretanto, por razões de clareza, a representação destas coordenadas será suprimida nas equações deste ponto em diante no texto, sendo que serão indicadas apenas quando for importante para a compreensão do tópico proposto.

$$I_{D(k)} = I_{J(k)}e^{-gd_{(k)}} + (1 - e^{-gd_{(k)}})\Lambda_{(k)}, \tag{4.5}$$

onde $I_{D(k)}$ é a imagem referência com degradação, $I_{J(k)}$ é a imagem de referência sem degradação, $gd_{(k)}$ é o parâmetro análogo ao produto atenuação-distância do modelo físico e $\Lambda_{(k)}$ é o parâmetro análogo à da luz ambiente, com $k \in \{R, G, B\}$.

Esta representação é utilizada na geração das imagens I_{cd} e I_{ce}, resultantes dos blocos de reforço e atenuação de cor e contraste na Fig. 4.1, respectivamente. Na forma dada na Eq. 4.5, a imagem resultante apresentará degradação acentuada em relação a imagem de referência. Tanto $I_{D(k)}$ como $I_{J(k)}$ podem ser obtidas como a imagem resultante da Eq. 4.5, desde que uma seja conhecida e a outra a que será gerada. Neste contexto, I_{cd} é determinada assumindo-se $I_{D(k)} = I_{cd(k)}$ e $I_{J(k)} = I_{hsr(k)}$ na Eq. 4.5, obtendo-se:

$$I_{cd(k)} = I_{hsr(k)}e^{-gd_{(k)}} + (1 - e^{-gd_{(k)}})\Lambda_{(k)}. \tag{4.6}$$

Como resultado do módulo de reforço de cor e contraste, a imagem I_{ce} deve possuir características de contraste e cor melhores que I_{hst}. Assim, adota-se I_{hst} como referência com de-

gradação e I_{ce} como referência sem degradação na Eq. 4.5, ou seja, $I_{D(k)} = I_{hst(k)}$ e $I_{J(k)} = I_{ce(k)}$, obtendo-se:

$$I_{ce(k)} = \frac{I_{hst(k)} - (1 - e^{-gd_{(k)}})\Lambda_{(k)}}{e^{-gd_{(k)}}}. \tag{4.7}$$

Neste trabalho, a intensidade dos pixels das imagens é representada dentro do intervalo $[0, 1]$ (normalização) em cada canal de cor. Imagens subaquáticas tendem a apresentar faixa dinâmica menor que a unidade, causada pela presença da água. Assim, a cor preta, cuja representação seria o pixel com valores $(0, 0, 0)$, tende a não ocorrer em uma cena na água, sendo que o valor mais próximo indicado pela coordenadas $(\min(R), \min(G), \min(B))$. Similar interpretação pode ser aplicada à cor branca, representada pelas coordenadas $(1, 1, 1)$ no espaço de cores. Com a presença da água o valor mais próximo é dado por $(\max(R), \max(G), \max(B))$. Mesmo imagens com cena em ambiente de água clara e limpa, podem apresentar faixa dinâmica menor que a unidade. Assim, pode-se inferir que quanto menor a turbidez e degradação imposta pela água, mais próxima estará a faixa dinâmica de 1.0 para cada canal de cor. Ou seja, menor será a distância dos valores mínimo e máximo da imagem (em cada canal de cor) aos limites 0 e 1, respectivamente. A partir desta premissa, estima-se que a diferença entre o valor por canal de cor de um dado pixel aos valores máximo e mínimo da imagem, guardem relação com o produto atenuação-distância do modelo físico. Em outras palavras, assume-se que quanto maior a diferença entre o valores de um pixel aos valores máximo e mínimo, maior será a distância física percebida do objeto ou ponto da cena em relação ao sensor (câmera).

Esta concepção é considerada na representação na Eq. 4.5 pelo índice $gd_{(k)}$, cuja expressão é indicada na Eq. 4.8. O objetivo é obter uma condição análoga ao produto entre índice de atenuação da água e a distância do objeto ao sensor, definida no modelo físico. Entretanto, como são estabelecidas duas referências para a distância, ou seja, os limites 0 e 1, se utiliza uma combinação simultânea das diferenças entre o pixel e os valores máximo e mínimo por canal de cor para a analogia com a distância. Esta combinação é ponderada através média geométrica entre as diferenças, conforme indicado na Eq. 4.8. Na interpretação proposta, considera-se que o sensor (câmera) esteja duplamente posicionado no espaço de cores, especificamente, nos pontos $(0, 0, 0)$ e $(1, 1, 1)$. Em consequência, há necessidade de realizar-se a correção da posição da cena em relação à câmera. O ajuste da posição da câmera é realizado pelo fator $r_{b(k)}$, que corresponde ao segmento da faixa dinâmica não ocupado devido à degradação, normalizado à faixa dinâmica.

$$gd_{(k)} = r_{b(k)}\sqrt{[\max(I_{(k)}) - I_{(k)}][I_{r(k)} - \min(I_{r(k)}]}, \tag{4.8}$$

onde $I_{(k)}$ é a imagem de entrada. O fator $r_{b(k)}$ estima a turbidez devido à água. É definida pela relação entre a redução experimentada pela faixa dinâmica (em relação ao valor ideal 1) e seu valor efetivo. Sendo expresso por:

$$r_{bw(k)} = \frac{1 - b_{w(k)}}{b_{w(k)}}. \tag{4.9}$$

Para ilustrar a relação de $gd_{(k)}$ com a imagem, são indicadas na Fig. 4.3 duas imagens subaquáticas com cenas apresentando níveis bastante distintos de degradação. a Fig.4.3(a) mostra imagem com baixo nível de degradação, enquanto a Fig.4.3(b) apresenta imagem com turbidez moderada-alta e *color cast* verde-azulado. Em conjunto, são incluídos os *arrays* correspondentes ao parâmetro $g_{d(k)}$ e suas respectivas componentes no espaço de cores RGB, $(g_{d(R)}, g_{d(G)}, g_{d(B)})$. Também são indicados os parâmetros máximo, mínimo e faixa dinâmica por canal de cor para cada imagem de entrada.

Comparando as representações visuais obtidas para os parâmetros $g_{d(k)}$ para ambas as imagens, verifica-se que para a imagem de entrada na Fig. 4.3(a) com baixo nível de degradação, o *array* de $g_{d(k)}$ mostra valores muito baixos (tendendo ao preto), com leve tonalidade verde. Esta tonalidade para $gd_{(k)}$) indica que o canal verde apresenta maior distorção, embora esta seja bastante leve em todos os canais. Isto pode ser verificado também, através da representação das respectivas componentes $g_{d(R)}$, $g_{d(G)}$ e $g_{d(B)}$. Percebe-se que o campo de pixels para o canal verde com maior intensidade em relação aos demais. Corroborando, a faixa dinâmica do canal verde é mais estreita que os demais, também sugerindo maior degradação para este canal. Para a imagem da Fig. 4.3(b), a apresentação visual do *array* de $g_{d(k)}$ mostra predominância do canal vermelho, sugerindo que este canal é o mais afetado pela degradação na imagem de entrada. Este condição também pode ser observada tanto pela menor intensidade nos pixels nos *arrays* das componentes verde e azul, como pela faixa dinâmica. O canal vermelho mostra faixa dinâmica sensivelmente mais estreita que a dos demais canais, também indicando maior degradação.

O parâmetro $\Lambda_{(k)}$ faz analogia à luz difundida no ambiente no modelo físico. No método proposto, $\Lambda_{(k)}$ é determinado a partir do valor máximo para $gd_{(k)}$, ponto do espaço de cores onde a imagem apresenta maior degradação. Ou seja,

$$\frac{\partial gd_{(k)}}{\partial I_{(k)}} = 0. \tag{4.10}$$

Realizando-se a diferenciação sobre a Eq. 4.8, e $\Lambda_{(k)}$ substituindo a imagem $I_{(k)}$, obtém-se:

$$\Lambda_{(k)} = \frac{\max(I_{(k)}) + \min(I_{(k)})}{2}. \tag{4.11}$$

O valor de $\Lambda_{(k)}$ corresponde ao valor central da faixa dinâmica da imagem.

A Fig. 4.4 mostra o resultado para o ajuste de cor e contraste proporcionado pela representação discutida acima. Utiliza-se a mesma imagem da Fig. 4.2 para facilitar a percepção visual do efeito produzido com as Eqs. 4.6 e 4.7. A imagem da Fig. 4.4(a2) mostra redução no efeito da turbidez e maior intensidade nas cores, com melhora na percepção do contraste. O histograma da Fig. 4.4(b2) reflete estas considerações, mostrando melhor distribuição da intensidade dos pixels no interior de cada canal de cor e entre eles. A imagem 4.4(a4) mostra o efeito

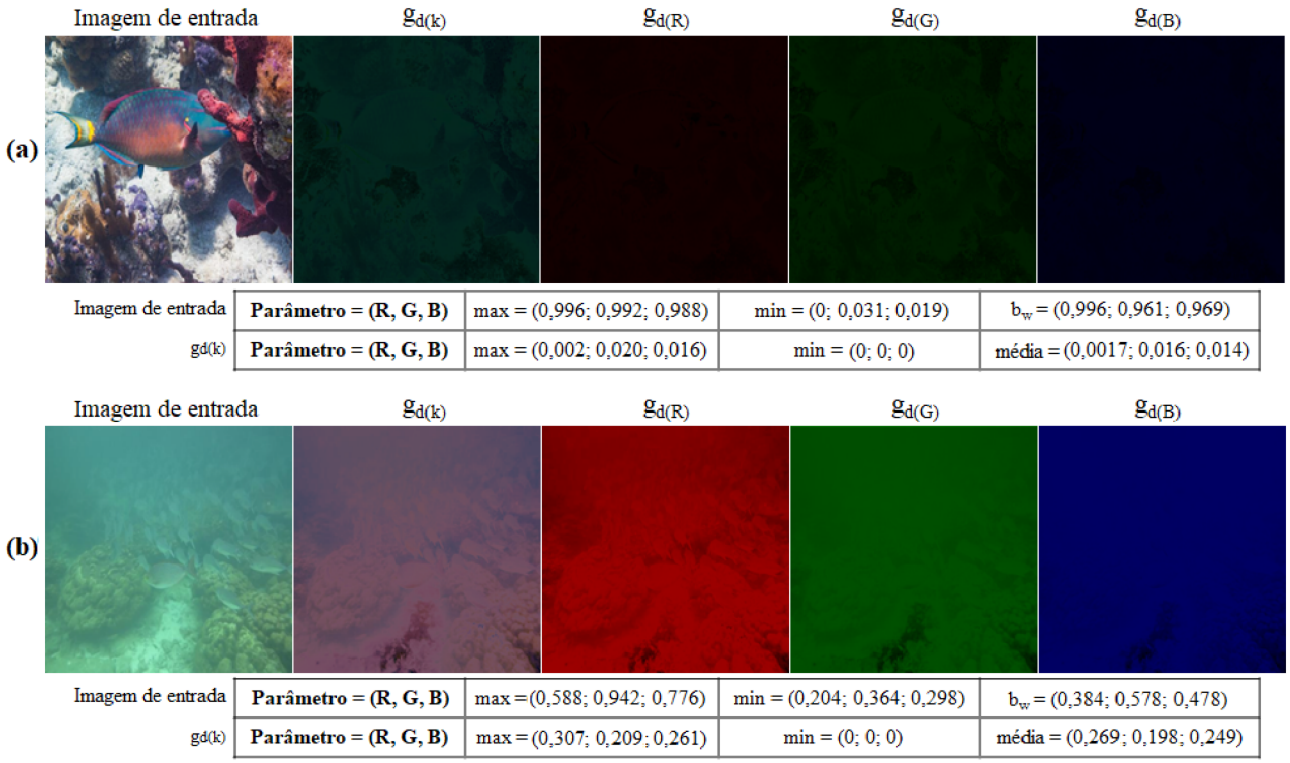


Figura 4.3: Exemplos de imagens subaquáticas mostrando cenas com diferentes níveis de degradação. (a) Imagem com baixo nível de degradação, (b) Imagem com degradação moderada-alta e *color cast* verde-azulado.

de atenuação obtido por meio de 4.6. Há o aumento na percepção da turbidez e borramento, com atenuação nas cores e no contraste. De maneira similar ao reforço, o histograma da Fig. 4.4(b4) reflete a percepção da cena com maior degradação, através da redução na intensidade dos pixels por canal de cor e maior desequilíbrio entre eles. O canal vermelho fica restrito à região de baixas intensidade, enquanto o azul e o verde assumem as regiões de média e alta intensidade, respectivamente. Interessante notar a influência do *color cast* na degradação. O canal dominante na imagem de entrada é o verde, conforme indicado pelos valores da intensidade média dos pixels na Fig. 4.4. Esta predominância é devido a tonalidade do *color cast*, verde-azulado, favorecendo ambos os canais verde e azul, mas com favorecimento ao verde.

4.4 Preparando o realce de textura

O borramento (*blurring*) provocado pela turbidez da água gera perda de definição de bordas e contornos na cena, reduzindo percepção de aspectos estruturais. A textura está relacionada à estrutura, mas fortemente vinculada a detalhes e contornos pertencentes à superfície de objetos, Gómez-Ríos et al. (2019). O efeito da turbidez é proporcional à distância entre o ponto da cena e a câmera. Em imagens com cenas amplas e objetos distantes da câmera, as dimensões de detalhes estruturais relacionados à textura são reduzidas e de difícil percepção. A percepção de detalhes associados a maior distância é naturalmente difícil devido a ação da água, mesmo para imagens com reduzida distorção. Assim, estratégias que minimizem este efeito através do

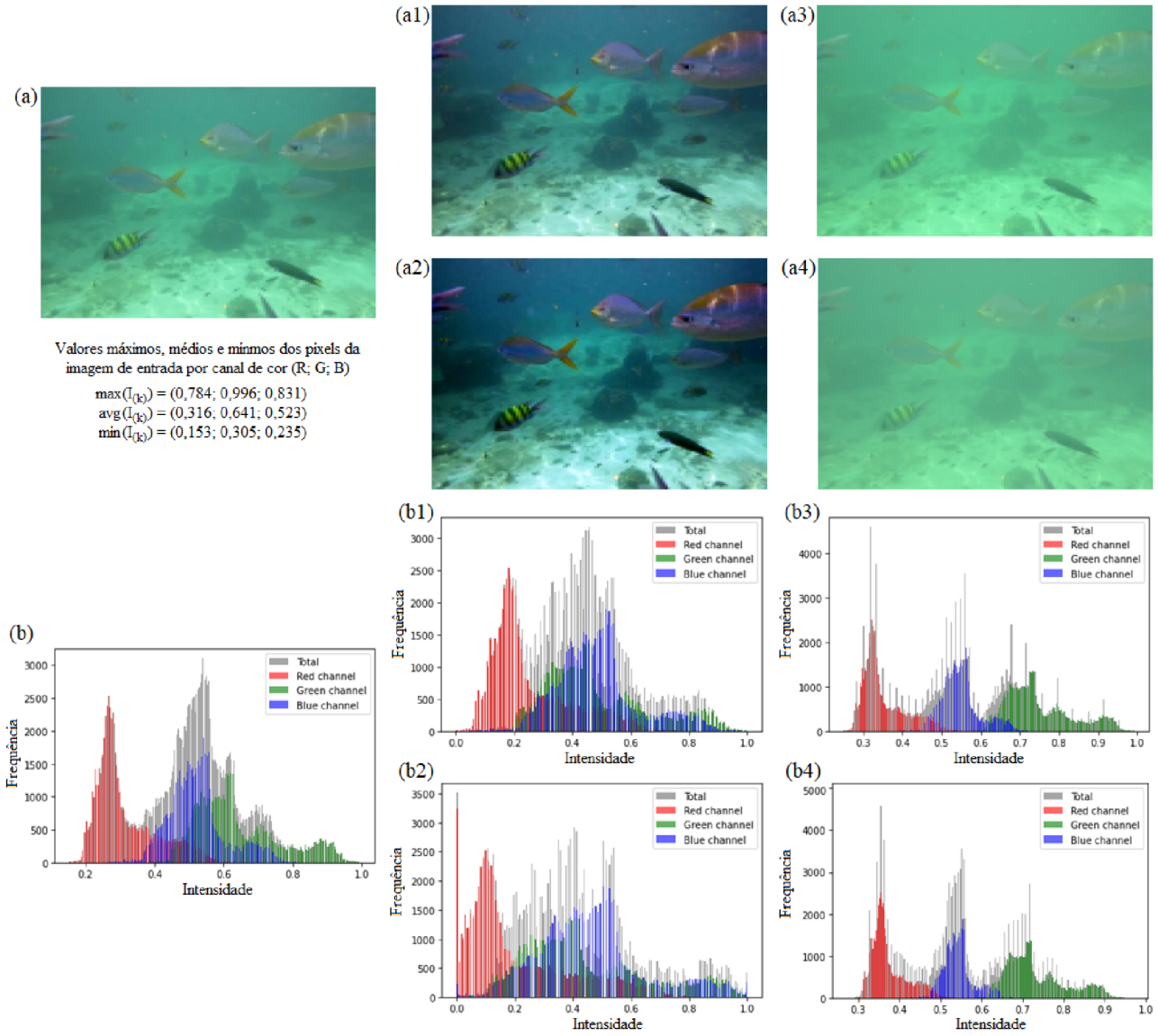


Figura 4.4: Imagem subaquática e versões resultantes das etapas de alteração no histograma e ajuste de cor e contraste. (a) Imagem de entrada $I_{(k)}$, (a1) Versão com histograma expandido $I_{hst(k)}$, (a2) Versão com reforço de cor e contraste $I_{ce(k)}$ (Eq. 4.7), (a3) Versão com histograma comprimido $I_{hsr(k)}$ e (a4) Versão com atenuação de cor e contraste $I_{cd(k)}$ (Eq. 4.6). (b) Histograma da imagem de entrada, (b1) Histograma de (a1), (b2) Histograma de (a2), (b3) Histograma de (a3), (b4) Histograma de (a4).

resgate visual de regiões percebidas distantes da câmera, tem impacto importante na qualidade e contexto da aplicação como inspeção e monitoramento, por exemplo.

A última etapa de cada *pipeline* do MD é dedicada a recuperação de textura. Nos módulos de reforço e redução de textura é utilizada filtragem gaussiana para a geração das imagens $I_{e(k)}$ e $I_{d(k)}$.

A imagem-guia para realce $I_{e(k)}$ deve apresentar reforço nas *features* associadas à textura. Essencialmente, o aspecto estrutural da imagem deve ser reforçado. Para a obtenção de $I_{e(k)}$, gera-se uma imagem contando majoritariamente, com bordas e contornos de $I_{ce(k)}$, através da diferença entre esta imagem e sua versão com filtragem gaussiana, ou seja:

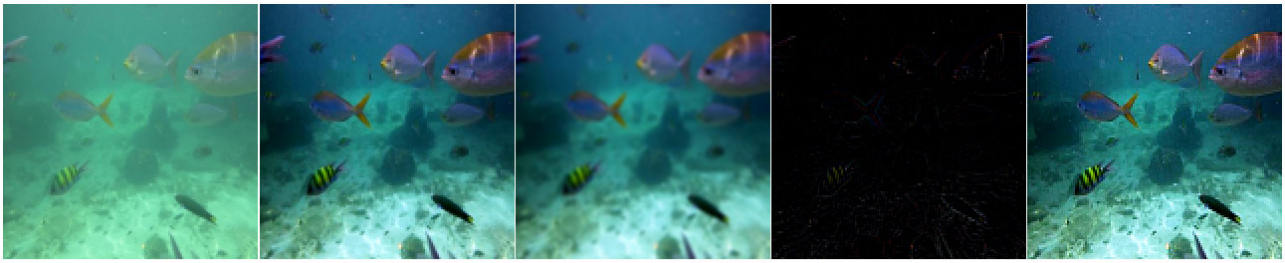


Figura 4.5: Conjunto de imagens utilizadas pelo bloco de reforço de textura do MD. (a)Imagem de entrada $I_{(k)}$, (b) Imagem resultante do bloco de reforço de cor e contraste $I_{ce(k)}$, (c) Imagem resultante da filtragem gaussiana $\mathcal{G}(I_{ce(k)})$, (d) Imagem contendo apenas elementos de alta-frequência $I_{L(k)}$ e (e) Imagem guia para realce $I_{e(k)}$.

$$I_{L(k)} = I_{ce(k)} - \mathcal{G}(I_{ce(k)}). \tag{4.12}$$

A operação descrita pela Eq. 4.12 resulta em uma imagem $I_{L(k)}$ contendo apenas elementos de alta frequência da imagem original. Assim, $I_{e(k)}$ é determinada por:

$$I_{e(k)} = I_{ce(k)} + I_{L(k)}. \tag{4.13}$$

Usando a Eq. 4.12 em 4.13, obtém-se:

$$I_{e(k)} = 2I_{ce(k)} - \mathcal{G}(I_{ce(k)}). \tag{4.14}$$

A imagem-guia para degradação $I_{d(k)}$ deve mostrar redução na percepção de bordas, contornos e texturas. Para a determinação de $I_{d(k)}$ utiliza-se a operação inversa descrita pela Eq. 4.14, considerando $I_{cd(k)}$ como entrada, obtendo-se:

$$I_{d(k)} = \frac{I_{cd(k)} + \mathcal{G}(I_{cd(k)})}{2}. \tag{4.15}$$

A imagem resultante da função de degradação $\hat{I}^*_{(k)}$, que substituirá $\hat{I}_{(k)}$ na função de perda, é obtida por:

$$\hat{I}^*_{(k)} = \hat{I}_{(k)} + I_{d(k)} - I_{e(k)}. \tag{4.16}$$

Utilizando-se a imagem da Fig. 4.2 para efeito de comparação, a Fig. 4.6 mostra as versões envolvidas nos procedimentos descritos pelas Eqs. 4.12 e 4.14 para obtenção de $I_{e(k)}$. Do modo similar, a Fig. 4.5 mostra as versões utilizadas na Eq. 4.15 na geração de $I_{d(k)}$.

Finalizando, apresenta-se na Fig. 4.7 as imagens de entrada e imagens-guia, juntamente com as respectivas distribuições dos pixels no espaço de cores RGB. O número de pontos gerados pela imagem no espaço de cores, indica o quanto diversificada se apresenta a distribuição de cores (tonalidades). Assim, indica-se o Número de Pontos Discretizados (NPD) gerados por cada imagem no espaço de cores. Embora não possa ser utilizado como critério formal de

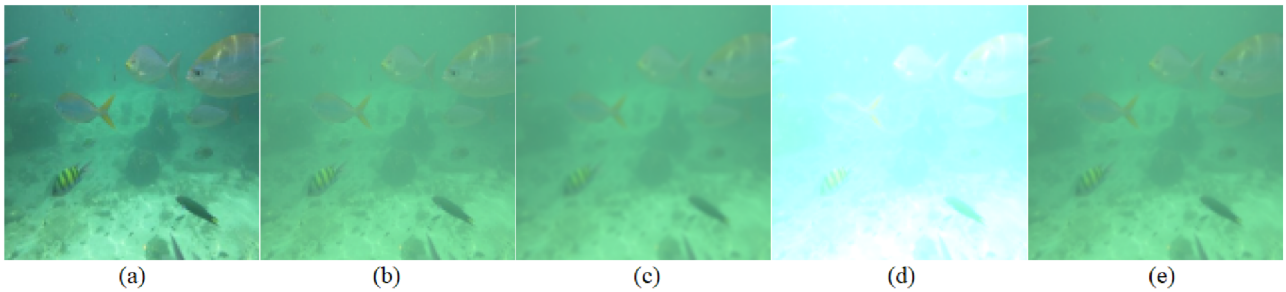


Figura 4.6: Conjunto de imagens utilizadas pelo bloco de atenuação de textura do MD. (a)Imagem de entrada $I_{(k)}$, (b) Imagem resultante do bloco de atenuação de cor e contraste $I_{cd(k)}$, (c) Imagem resultante da filtragem gaussiana $\mathcal{G}(I_{cd(k)})$, (d) Imagem com redução nos elementos de alta-frequência (numerador na Eq. 4.15) e (e) Imagem guia para degradação $I_{d(k)}$.

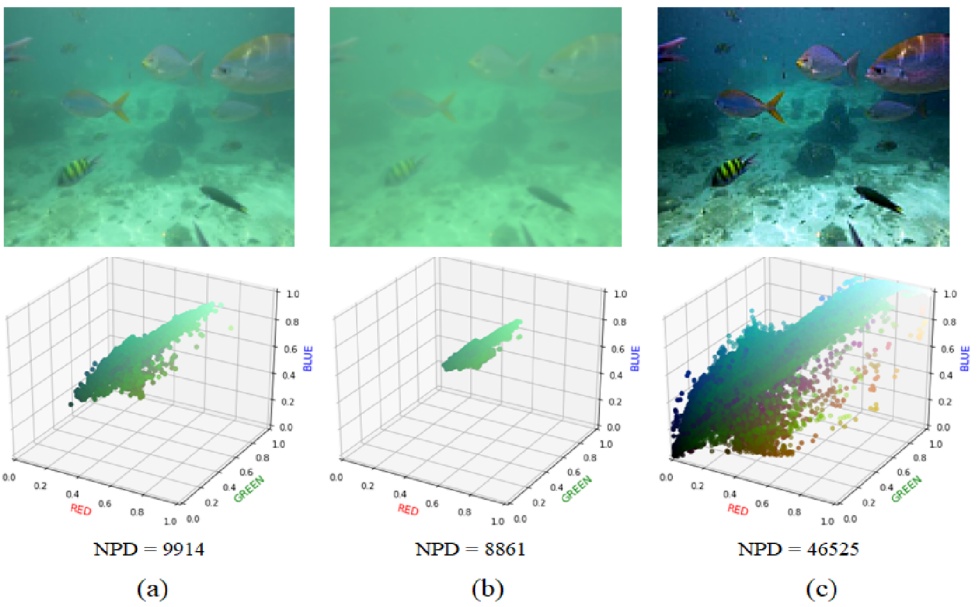


Figura 4.7: Exemplo de imagem de entrada, imagens-guia e respectivas distribuições no espaço de cores e NPD. (a)Imagem de entrada $I_{(k)}$, (b) Imagem $I_{d(k)}$, (c) Imagem $I_{e(k)}$

avaliação, em condições bastante restritas o NPD pode subsidiar a avaliação da qualidade da imagem. Para referência, todas as imagens possuem as dimensões $256 \times 256 \times 3$.

A percepção visual oferecida pela Fig.4.7 salienta a relação degradação-qualidade através do volume desenvolvido pelas imagens no espaço RGB e, implicitamente, pelas variações nas faixas dinâmicas. Na situação limite, em que a turbidez seja intensa e uniforme o suficiente de modo a não permitir a visualização de qualquer objeto na cena, a distribuição no espaço de cores tenderia a apresentar apenas um ponto. No sentido contrário, a imagem descrevendo amplo volume no espaço de cores indica maior diversidade nas cores, contraste e melhor qualidade.

4.4.1 Limitações ao tratamento de imagens

A expansão do histograma melhora o contraste global e o realce de cores, permitindo uma percepção mais natural da cena em imagens subaquáticas. No entanto, esta operação pode

produzir ou facilitar o surgimento de áreas saturadas ou de alta-intensidade na imagem. A situação ocorre principalmente em imagens com acentuado desequilíbrio entre os canais de cores, que apresentam regiões em que a intensidade dos pixels é muito superior a média do canal, denominadas de regiões *outlier* neste trabalho. Regiões *outlier* são geradas pela distribuição irregular da luz ambiente e pela interação desta com a turbidez e *color cast*.

As imagens com desequilíbrio entre os canais mostram valores mínimos muito reduzidos em pelo menos um dos canais de cores. Durante a expansão do histograma, o canal com menor intensidade dos pixels (desequilibrado) tende a produzir uma estreita faixa dinâmica. Em consequência, há uma forte compensação deste canal na normalização, produzindo valores elevados no valor associado ao canal no pixel. Se a imagem contém regiões *outlier*, tenderá a gerar áreas com saturação no imagem realçada. A Fig. 4.8 ilustra o caso descrito. A imagem mostrada apresenta desequilíbrio entre os canais, com o canal vermelho bastante atenuado em relação aos demais. Na versão com histograma expandido, indicada na Fig. 4.8(b), são geradas regiões saturadas e de alta intensidade, indicadas pelas caixas em vermelho. O equivalente ocorre na etapa de contração do histograma, onde estas regiões são atenuadas na versão da imagem com histograma contraído. Estas regiões são então propagadas ao longo dos *pipelines*, sendo realçadas na imagem de saída, mostrada na Fig. 4.8(c).

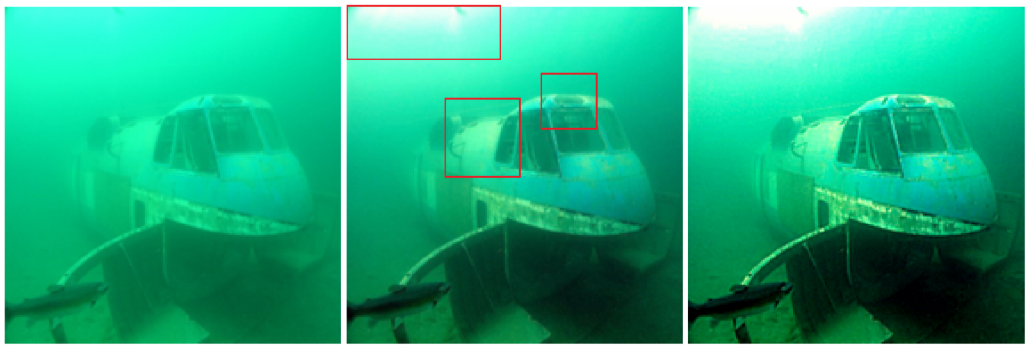


Figura 4.8: Exemplo de imagem com acentuado desequilíbrio entre os canais de cores. (a) Imagem original, (b) Imagem com expansão do histograma e (c) Imagem realçada de saída.

A estratégia para redução destas limitações segue a linha conceitual do trabalho, onde a informação associada a estas áreas saturadas e de alta intensidade é isolada e utilizada como parcela adicional na função de degradação. A expectativa é que a rede neural aprenda a compensar este incremento na degradação e reduza a intensidade destas áreas na imagem realçada. A Eq. 4.17 sintetiza a concepção descrita através da função de degradação, onde $I_{at(k)}$ corresponde à imagem contendo apenas a informação associada às áreas de alta intensidade.

$$\hat{I}^*_{(k)} = \hat{I}_{(k)} + I^*_{d(k)} - I_{e(k)} + I_{at(k)}. \tag{4.17}$$

Um mecanismo de atenção (*attention*) à estas áreas realiza o mapeamento da imagem de entrada, gerando a imagem $I_{at(k)}$ que contém as regiões *outlier*. Estas regiões são provocadas pela presença de pixels com alta intensidade e acentuado desequilíbrio entre os canais.

A determinação de $I_{at(k)}$ é realizada conforme a Eq. 4.18. Consiste na expansão do histograma da imagem $I_{a(k)}$, obtida através da Eq. 4.18, como segue:

$$I_{at(k)} = \frac{I_{a(k)} - min(I_{a(k)})}{max(I_{a(k)}) - min(I_{a(k)})}, \tag{4.18}$$

onde $I_{a(k)}$ contém a informação associada às áreas *outlier*, sendo determinada na Eq. 4.19.

$$I_{a(k)} = g_{r(k)} \mid av(I_{(k)}) - av_g(I_{(k)}) \mid \frac{I_{x(k)}}{I_{(k)}}, \tag{4.19}$$

onde $g_{r(k)}$ é um fator de reforço, calculado na Eq. 4.20, $av(I_{(k)})$ e $av_g(I_{(k)})$ são o valor médio por canal de cor da imagem de entrada $I_{(k)}$ e o valor médio global da imagem, respectivamente.

$$g_{r(k)} = 1 + \frac{thr_{(k)}}{av(I_{(k)})}. \tag{4.20}$$

A detecção do desequilíbrio nos canais de cores é executada pelo termo $\mid av(I_{(k)}) - av_g(I_{(k)}) \mid$. A imagem $I_{x(k)}$ contém os pixels com alta intensidade mapeados de $I_{(k)}$, por meio da Eq. 4.21. Pixels acima do limiar estabelecido por $max(I_{(k)})\frac{thr_{(k)}}{av(I_{(k)})}$ são considerados *outliers* e selecionados para compor $I_{x(k)}$.

$$I_{x(k)} = \begin{cases} I_{(k)} - max(I_{(k)})(1 - \frac{thr_{(k)}}{av(I_{(k)})}), & I_{x(k)} \geqslant 0 \\ 0, & I_{x(k)} < 0. \end{cases} \tag{4.21}$$

O termo $thr_{(k)}$ é descrito na Eq. 4.22 e corresponde ao valor limiar para seleção de um pixel como *outlier*. A investigação das características de imagens com regiões de alta intensidade mostrou que a diferença entre a mediana ($mdn_{(k)}$) e a média aumenta na presença de *outlier* na imagem de entrada. A detecção destes valores explora o fato de que a média é fortemente afetada pela presença deles na imagem de entrada, sendo assim utilizada para a seleção dos pixels.

$$thr_{(k)} = \begin{cases} av(I_{(k)}) - mdn(I_{(k)}), & thr_{(k)} \geqslant 0 \\ 0, & thr_{(k)} < 0. \end{cases} \tag{4.22}$$

5 METODOLOGIA PARA MELHORIA DA QUALIDADE DE VÍDEO SUBAQUÁTICO

É interessante a constatação da relativa carência de propostas na bibliografia voltadas para a melhoria de vídeos subaquáticos. Isto contrasta com o grande número de atividades submersas que são realizadas com a aquisição de sequências de vídeo. Atualmente, a aquisição de vídeos é facilitada não apenas pela disponibilização de câmeras à prova d'água, de baixo custo e grande popularidade, mas pela profusão de veículos robóticos operados remotamente ou autônomos e empregados, majoritariamente, em aplicações envolvendo manutenção, inspeção e exploração marinhas.

Neste capítulo apresenta-se a abordagem metodológica para a extensão do método aplicado a vídeo subaquático. Inicialmente, discute-se características e requisitos envolvidos no contexto da melhoria de vídeo e a relação com o método proposto. A seguir, apresenta-se a adequação proposta à metodologia.

5.1 Desafios à melhoria de vídeo

A principal diferença entre o tratamento de vídeo e o de imagem é a necessidade de garantir que o resultado para a ação de melhoria na qualidade dos *frames* do vídeo seja uniforme, coerente ao longo tempo. Em outras palavras, há a necessidade que o resultado apresente consistência temporal.

Na bibliografia, a opção mais empregada para o processamento de vídeo visando melhoria da qualidade é o realce ou restauração quadro-a-quadro, onde cada *frame* é tratado como uma imagem. Entretanto, variações na ação de realce de um *frame* para outro são percebidos como oscilações repentinas na cor e luz ambiente na cena como um todo ou em regiões desta. Estes efeitos, identificados como *flickering*, indicam limitação no desempenho de um dado método de realce.

Nas abordagens baseadas em *deep learning*, a consistência temporal depende do método de realce e da natureza do ambiente subaquático. O movimento da câmera e/ou objetos no ambiente subaquático durante a aquisição de uma sequência, estabelece um cenário desafiador, pois cenas diferentes e sucessivas são apresentadas ao método sob uma mesma condição ambiental (turbidez, *color cast*, luz ambiente). Em consequência, é esperado que o método apresente a mesma performance de realce em toda a sequência. Esta condição também implica em complexidade adicional, pois sequências subaquáticas são adquiridas sob condições ambientais constantes. Estas condições não se alteram durante a aquisição de *frames* sucessivos. Apesar da extração de *features* associadas à dinâmica das cenas ser favorecida durante o treinamento da rede, o aprendizado gerado pelo mapeamento de *features* relacionadas ao ambiente é limitado pela aparente monotonia deste contexto. Em consequência, o aprendizado se estabelece em função do que a dinâmica dos quadros oferece, e tende a apresentar insensibilidade a *features*

do ambiente. Mesmo o modelo sendo treinado com imagens UW diversificadas, onde o aprendizado esteja associado a um conjunto maior de *features*. Na inferência para vídeo, o ambiente contribuirá em menor escala e o resultado do realce será condicionado pelo movimento dos objetos e da câmera ao longo do vídeo. Em vídeos UW, este movimento apresenta variações muito pequenas (lentas) entre quadros sucessivos, compondo cenários muito semelhantes para o realce quadro-a-quadro, que predispõe o modelo a reagir mais facilmente para pequenas mudanças. Ou seja, o modelo torna-se mais sensível à mudanças no contexto da sequência. Isto conduz a resultados distintos no realce entre *frames* sucessivos, gerando *flickering*.

As questões discutidas acima impõem limitações no uso direto do algoritmo para realce de imagens subaquáticas apresentado no capítulo anterior. As causas destas limitações são discutidas nas seções a seguir, assim como a alternativa desenvolvida para a extensão do método para o realce de vídeo.

5.2 Alterações no histograma e *flickering*

O método para realce de imagens usa como informação inicial dos *pipelines* no módulo de degradação, imagens obtidas das operações de expansão e contração do histograma aplicadas sobre a imagem de entrada, como expresso nas Eqs. 4.1 e 4.2. Ambas as operações são definidas pelos valores máximo e mínimo por canal de cor entre pixels da imagem, sendo, portanto, o resultado dependentes destes valores.

Os *frames* sucessivos de uma sequência de vídeo apresentam um contexto dinâmico proposto pelo movimento da câmera e/ou dos objetos. Mesmo que as condições impostas pelo ambiente não mostrem variação significativa entre *frames* sucessivos, a dinâmica dos objetos e câmera pode gerar alterações expressivas em cenas sucessivas. Isto sugere de maneira consistente que os valores máximos é mínimos para cada *frame* são próprios e independentes entre si e que podem, efetivamente, variar entre quadros sucessivos.

À luz do método desenvolvido neste trabalho, estas considerações são significativamente importantes, em função de que as operações sobre o histograma da imagem (*frame*) dependem diretamente destes valores. Ocorrendo variações entre quadros sucessivos, implica em variações no resultado das operações sobre o histograma. Estas variações são propagadas ao longo dos *pipelines*, resultando em oscilações na ação de realce, ou seja, resultando em *flickering*. A Fig. 5.1 exemplifica a situação discutida acima. São mostradas dois exemplos com sequências de *frames* dos vídeos identificados por "Sonar"(Ryan (2021)) e "Mobula"(Guerra (2022)) e as respectivas versões com expansão do histograma. "Sonar"mostra ambiente com turbidez moderada-baixa e *color cast* levemente azulado e objetos próximos da câmera. "Mobula"apresenta cena ampla com objetos distantes da à câmera, com intenso *color cast* verde-azulado e turbidez elevada. Em ambos as sequências, pode-se perceber as diferenças nas imagens resultantes da expansão do histograma, com variações na cor dos objetos e intensidade ("Mobula") ou coloração da luz de fundo ("Sonar"). Estas variações (*flickering*) são facilmente percebidas durante a execução do vídeo realçado, indicando deficiência na consistência temporal.

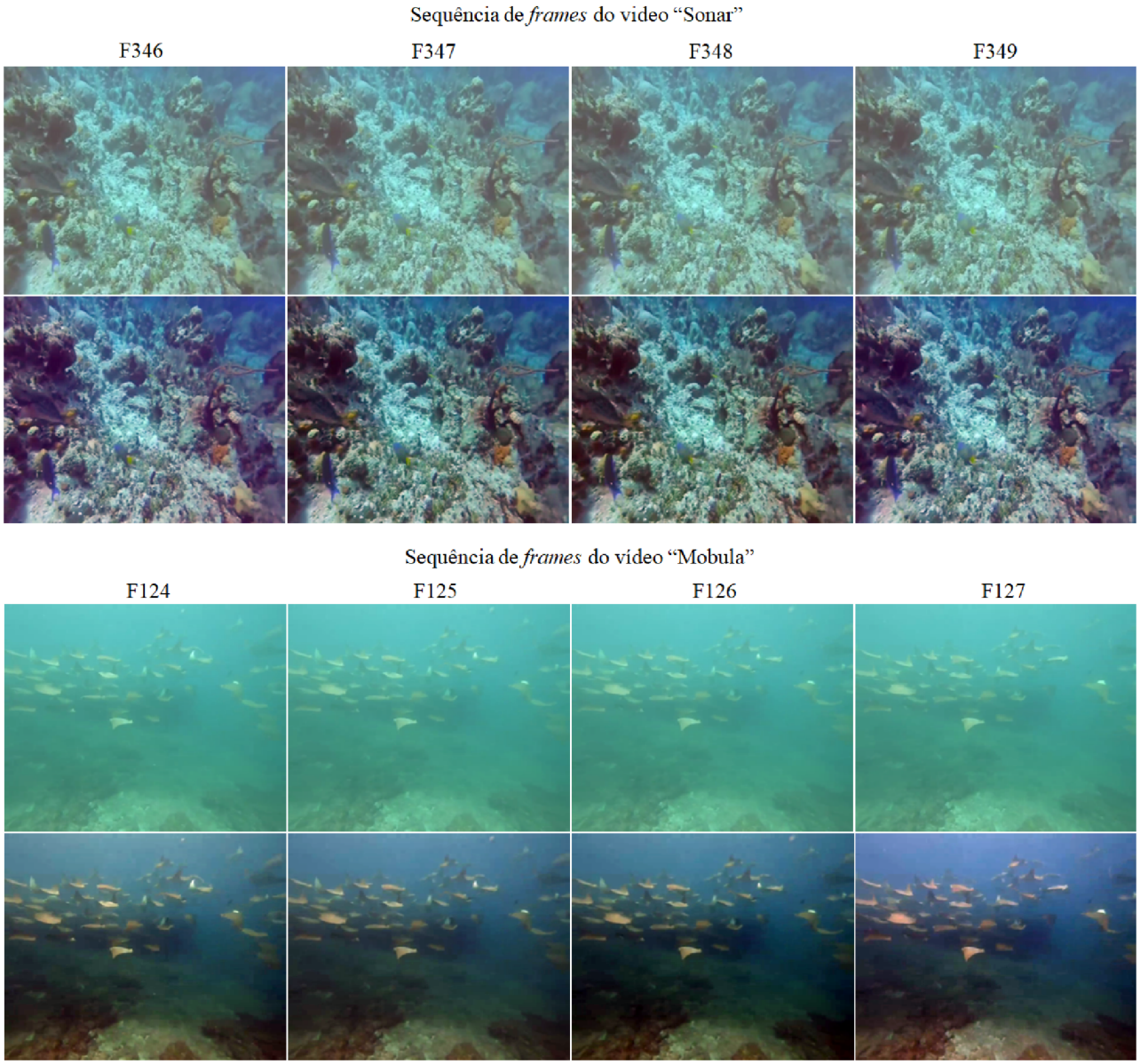


Figura 5.1: Sequências de *frames* dos vídeos "Sonar"e "Mobula"e respectivas versões com histograma expandido.

A Fig. 5.2 mostra a evolução no tempo para os valores máximos, mínimos, faixa dinâmica ($b_{w(k)}$) e média ($avg_{(k)}$) para os 400 primeiros *frames* de ambos os vídeos. Em conjunto, são mostradas os comportamentos das taxas de variação no tempo (derivadas) da faixa dinâmica e da média. Os retângulos em amarelo demarcam a região onde se encontram os *frames* indicados na Fig. 5.1. A observação das características dos valores máximos e mínimos, percebe-se oscilações bastante intensas em intervalos de tempo muito pequenos. Esta condição é apresentada também na evolução das taxas de variação no tempo das faixas dinâmicas $\frac{d(b_w)}{dt}$. As oscilações são predominantes nos mínimos para o vídeo "Sonar", com especial atenção para o canal azul. O vídeo "Mobula"apresenta maiores oscilações nos valores máximos, com destaque para o canal vermelho. Estas variações nos valores impacta sobre as faixas dinâmicas e, em consequência, sobre o resultado das operações com histograma.

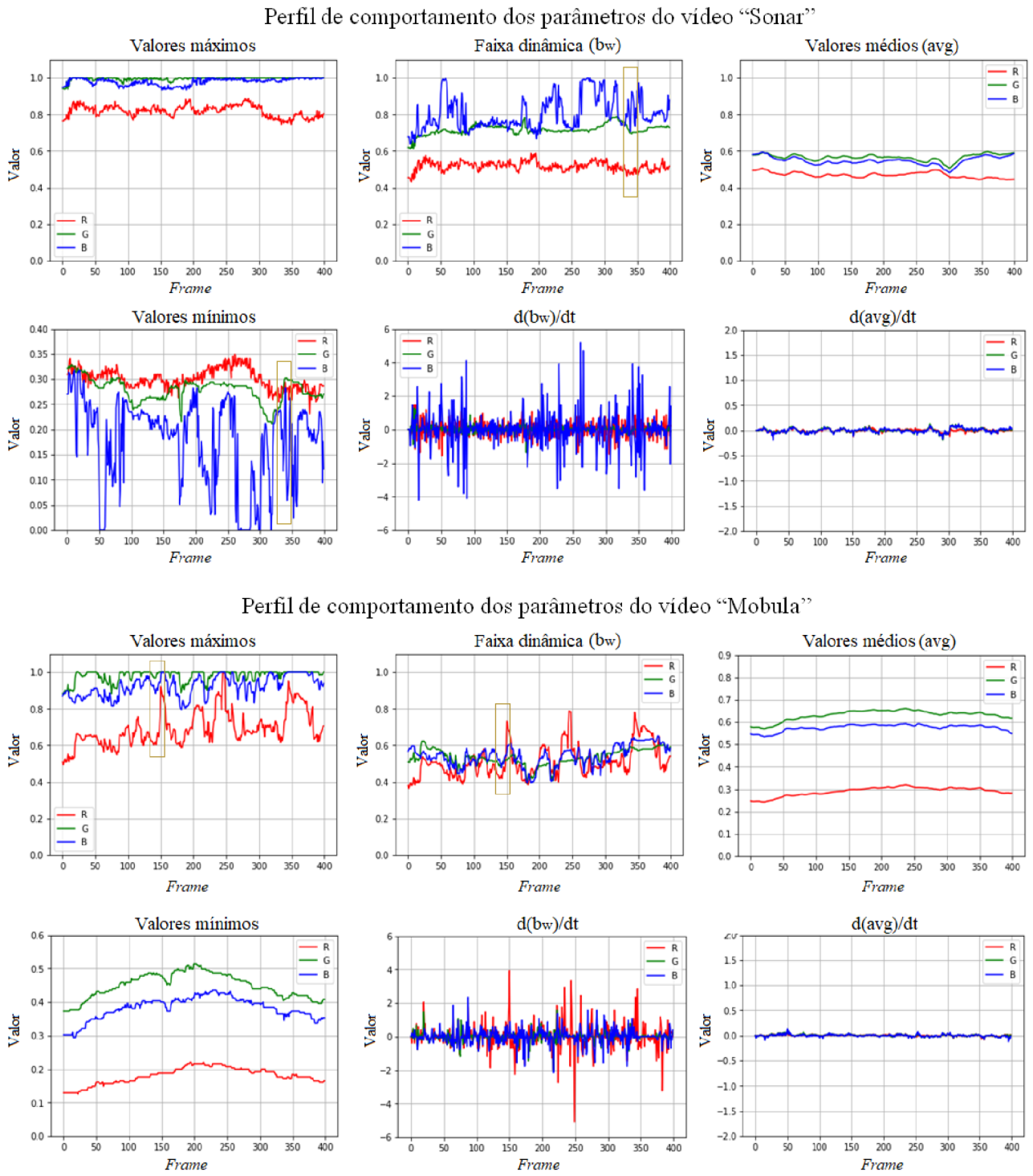


Figura 5.2: Comportamento no tempo dos valores máximo, mínimo, médio e faixa dinâmica para os vídeos "Sonar", Ryan (2021) e "Mobula", Guerra (2022).

Embora máximos e mínimos definam estas operações nos ajustes dos histogramas, os mesmos não são representativos do *frame* como um todo. O contexto físico-ambiental das cenas não varia significativamente uma sequência de vídeo UW contínua (sem cortes ou edições). Em outras palavras, o panorama definido pela turbidez e pelo *color cast* da água não mostra variações importantes ao longo de um vídeo UW obtido de forma contínua em um mesmo ambiente ou região submersa. Esta consideração torna-se mais evidente quando é considerada a evolução (intervalo de tempo) entre *frames* consecutivos. Isto pode ser inferido a partir do comportamento das médias, com variações muito mais lentas, sugerindo valores constantes entre as transições dos quadros. A característica da taxa de variação da média no tempo ($\frac{d(avg)}{dt}$) corrobora esta interpretação, com valores muito próximos de zero e bastante menores que os relacionados à faixa dinâmica ($\frac{d(b_w)}{dt}$). Como a média é representativa do conjunto de pixels, pode-se considerar que o contexto da cena não varia significativamente de um *frames* para outro.

A metodologia para realce de imagens UW apresentada traz limitação quando aplicada ao realce de vídeos subaquáticos, pois não garante consistência temporal devido à sensibilidade das operações de ajuste do histograma às variações de máximos e mínimos. Assim, elaborou-se estratégia para adequação do algoritmo proposto ao realce de vídeo, mantendo-se a concepção original proposta para imagem, que será apresentada nas seções à seguir.

5.3 Máximos e mínimos como *outliers*

Uma das considerações mais importantes na discussão anterior, está na variação "suave" do valor médio das intensidades dos pixels dos *frames* ao longo de um vídeo UW contínuo. Diferente dos valores máximo e mínimo, este não varia significativamente na transição entre *frames*. Como o valor médio é representativo do conjunto de pixels, pode-se inferir que uma quantidade considerável destes, não sofre variação apreciável de intensidade nas transições entre *frames* consecutivos. Considera-se esta hipótese plausível para vídeos subaquáticos contínuos, ou seja, vídeos sem cortes ou transições instantâneas de um contexto de cena para outro, como ocorre em vídeos editados. Isto é reforçado pelo fato de que cenas consecutivas em um vídeo contínuo, tendem a mostrar as mesmas condições ambientais, ou seja, mesmas condições de turbidez e *colorcast*. A percepção de cores, contraste e luz dispersa no ambiente tendem a ser muito semelhantes.

Assim, assume-se como premissa que uma ampla maioria dos pixels não apresenta variação significativa entre *frames* consecutivos. Esta consideração é balizada no comportamento da média, na constância dos fatores ambientais e na percepção de cor e contraste na cena. A partir disto, pode-se avaliar a relação entre a cena, através de seu valor médio, e os valores máximo e mínimo. Novamente, estes últimos não são representativos do conjunto, mas valores pontuais para cada canal de cor, e que variam entre *frames* conforme a dinâmica da cena e o movimento da câmera. Especificamente, o movimento dos objetos no vídeo UW produz situações em que as respectivas distâncias variam em relação a câmera. Em um dado instante,

podem surgir objetos muito próximos à câmera. Estes apresentam condições de visibilidade muito melhor que aqueles mais afastados. Isto implica que estes objetos, ou regiões deles, tenham melhor (ou muito melhor) percepção de cor. Na quantização, o sensor da câmera gera pixels com maior intensidade e equilíbrio entre os canais de cores. Nestes pixels, os efeitos da turbidez e *color cast* são fortemente reduzidos, gerando diferenças em relação aos demais pixels do *frame* (valores fora do padrão ou *outliers*). Efeito similar é provocado pelo movimento da câmera, que além de promover a variação de distâncias dos objetos à câmera, pode incluir a fonte de luz na transição entre cenas de *frames* sucessivos. Esta situação pode gerar regiões de alta intensidade na transição entre as cenas, resultando que os pixels nestas regiões tenham valores para os canais de cores muito discrepantes dos demais, gerando valores *outliers*. Vale destacar que a intensidade da fonte de luz quando presente na cena é, essencialmente, resultado da sua posição (hora do dia), turbidez da água e profundidade. Assim, o panorama dado pelo movimento de objetos e câmera favorece que valores máximos e mínimos possam assumir a condição de valores fora do padrão, valores *outliers*, na distribuição dos pixels no espaço de cores.

Para verificação desta condição, procedeu-se à determinação da média envolvendo os pixels localizados nas regiões extremas do volume do *frame* no espaço de cores. O comportamento destas médias *frame* a *frame* pode fornecer indicativo sobre o posicionamento de mínimos e máximos como *outliers*. Para tanto, gera-se um *frame* modificado, onde os pixels nas regiões extremas são substituídos pelo valor médio desta região por canal de cor. Uma vez que nestes *frames* modificados, a evolução no tempo para a faixa dinâmica mostra redução das variações durante transições e a que média global por canal de cor se mantém bastante próxima ao *frame* original, assume-se que máximos e mínimos como valores fora do padrão.

Como exemplo, nas Figs. 5.3 e 5.4 são apresentados o comportamento da média e da faixa dinâmica para os *frames* original e as versões corrigidas com a média calculada sobre os pixels superiores correspondentes a 1% e 3%, respectivamente aos vídeos "Sonar"e "Mobula". Dois pontos se destacam na comparação entre o comportamento dos vídeos originais e as respectivas versões para 1% e 3%. Primeiro, a grande similaridade entre os comportamentos dos valores médios, mostrando que a média do *frame* não é afetada significativamente nas modificações efetuadas. Segundo, a redução nas oscilações entre transições, bastante evidentes nos perfis da sequência original e atenuadas nas versões modificadas, percebido tanto no comportamento das faixas dinâmicas quanto na respectivas taxas de variação no tempo. As variações mostradas nos perfis corrigidos representam a evolução do conteúdo das cenas ao longo do tempo, com mudanças lentas de contexto em função do movimento da câmera ou dos objetos. As taxas de variação da faixa dinâmica também evidenciam a atenuação nas transições entre quadros consecutivos. A versão corrigida pela média dos 3% superiores mostra intensidade ligeiramente maior na atenuação nas oscilações.

Embora como exemplo foram apresentados apenas dois vídeos, o estudo foi realizado envolvendo um conjunto maior de sequências subaquáticas, com resultados similares ao exposto. Em consequência, as premissas que fundamentaram a exposição acima, foram adotadas na con-

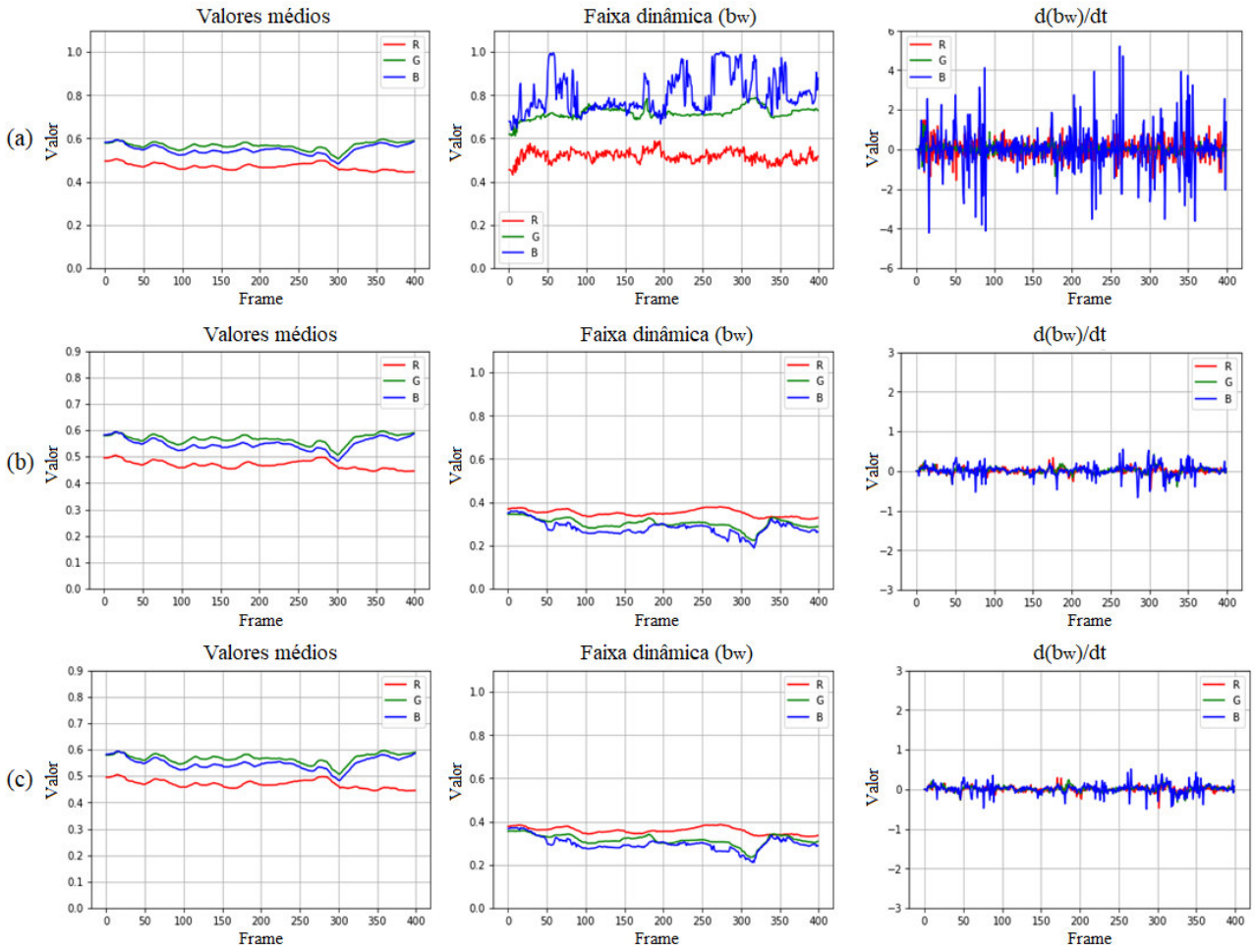


Figura 5.3: Comportamento no tempo dos valores médios e faixa dinâmica e taxa de variação da faixa dinâmica para o vídeo "Sonar". (a) Vídeo original, (b) Versão corrigida em 1% e (c) Versão corrigida em 3%.

cepção da estratégia de adequação do método para o realce de vídeo, discutida na próxima seção.

5.4 Módulo de degradação para realce de vídeo subaquático

A adequação da metodologia para a melhoria da qualidade de vídeo está orientada, essencialmente, para se prover consistência temporal à sequência resultante e realçada, mantendo-se a concepção original do método, ou seja, sem exigir dados anotados e utilizando a redução da degradação no *frame* como tarefa-pretexto para a condução do aprendizado da rede.

As causas da inconsistência no tempo na ação de realce do método para imagens, são as operações sobre o histograma. Como o resultado destas operações constitui a informação inicial para realce, variações na cor e/ou luz ambiente produzidas envolvendo quadros sucessivos são propagadas no *pipeline* de degradação, afetando o desempenho do modelo. Assim, a estratégia adotada para adequação do método consiste em substituir as operações de expansão e contração do histograma, por um algoritmo que gere versões do *frame* sem estas variações. Estas versões

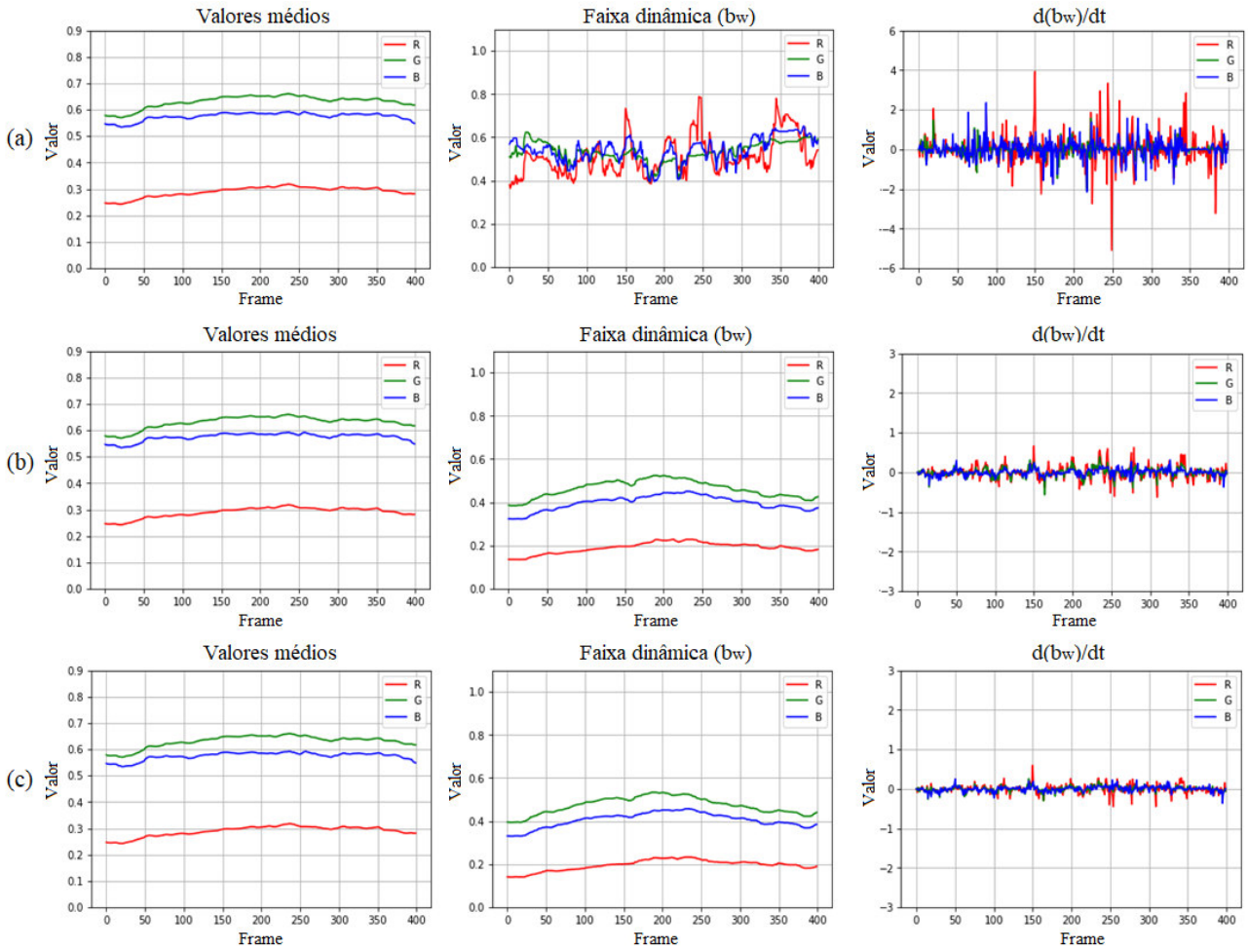


Figura 5.4: Comportamento no tempo dos valores médios e faixa dinâmica e taxa de variação da faixa dinâmica para o vídeo "Mobula". (a) Vídeo original, (b) Versão corrigida em 1% e (c) Versão corrigida em 3%.

passam a ser utilizadas como informação inicial para a metodologia proposta. A Fig. 5.5 apresenta o diagrama do módulo de degradação para vídeo, composto pelo Módulo de Geração dos *Frames* de Referência (GFR), que substitui as operações sobre o histograma da Fig. 4.1, e o Módulo de Geração dos *Frames*-Guias Duais (GFD).

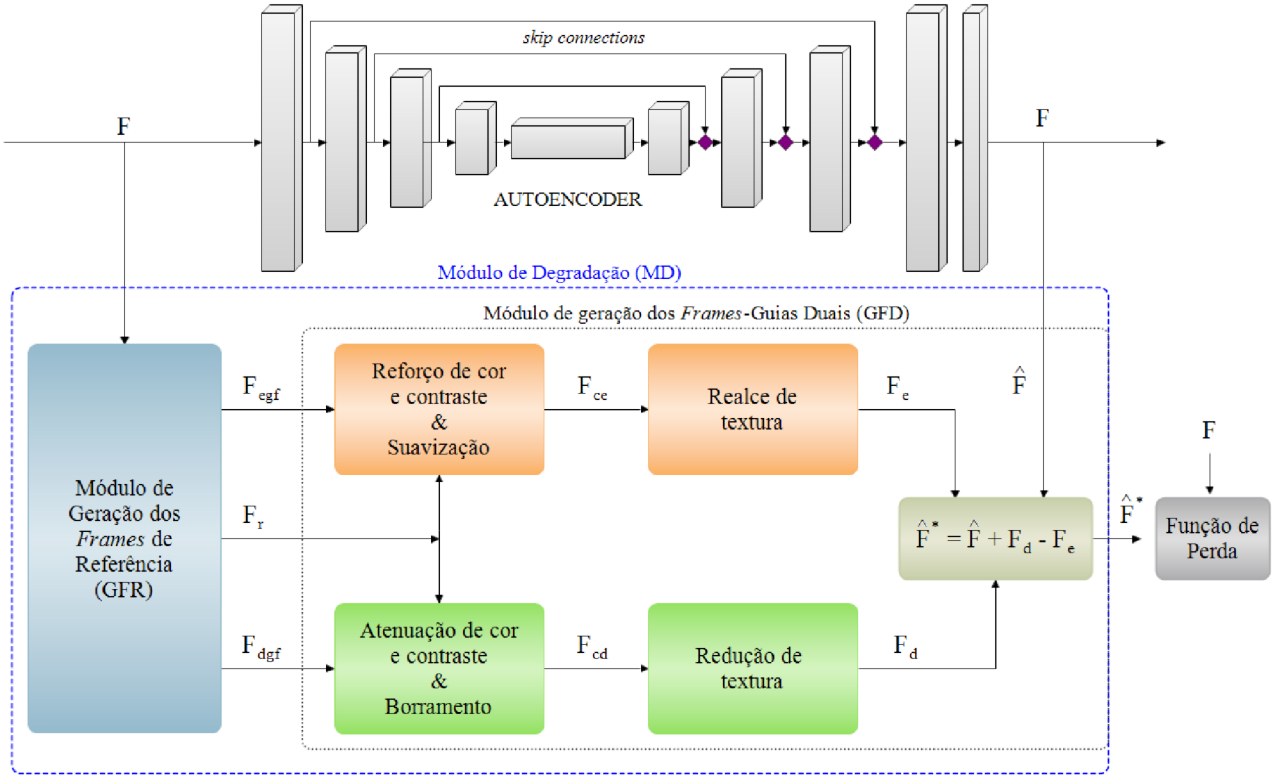


Figura 5.5: Diagrama do Módulo de Degradação para realce de vídeo

A estrutura é similar àquela concebida para realce de imagens, onde F e $\hat{F}$ são os *frames* de entrada e da saída decodificada pelo autoencoder, respectivamente. F_e é o *frame*-guia para realce e F_d o *Frame*-guia para degradação. $\hat{F}^*$ é o *frame* degradado resultante do MD. F_{ce} e F_{cd} são versões intermediárias para reforço e atenuação de cor e contraste, respectivamente. F_{egf}, F_{dgf} e F_r são versões de referência para realce, degradação e entrada, resultantes do GFR. Ambos os módulos são discutidos na seções a seguir.

5.4.1 Módulo de Geração dos Frames de Referência

A principal diferença em relação à estrutura para imagens é a substituição das operações sobre o histograma pelo módulo GFR. Este módulo gera informações de referência inicial para realce e degradação, F_{egf}, F_{dgf}. Em conjunto, gera versão do *frame* de entrada com valores máximos e mínimos corrigidos, através do método da média de valores extremos, discutida anteriormente. Esta versão é utilizada como fonte para determinação dos parâmetros necessários nas etapas de reforço e atenuação de cor e contraste. O diagrama do módulo GFR é apresentado na Fig. 5.6.

A média dos valores superiores e inferiores por canal de cor do *frame* de entrada, é utilizada

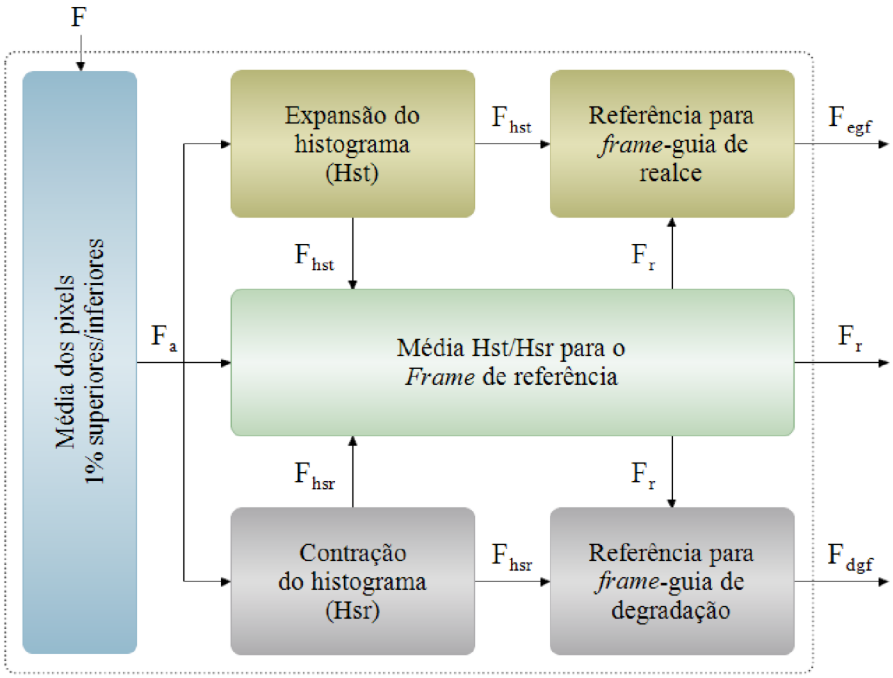


Figura 5.6: Diagrama do Módulo de Geração dos *Frames* de Referência (GFR)

para substituir os valores dos pixels nestas regiões. Adotou-se como *profundidade* de cálculo para a média o valor de 1%. Este valor foi selecionado a partir de avaliações experimentais, pois além de produzir mínima interferência na informação original, mostrou-se plenamente satisfatório para a maioria dos vídeos analisados no trabalho. Entretanto, casos particulares onde vídeos requerendo valores percentuais maiores para a média são discutidos em seção específica no capítulo destinado aos resultados do trabalho. O *frame* resultante da correção pela média, F_a, é utilizado na determinação das versões com histograma expandido ($F_{hst(k)}$) e contraído ($F_{hsr(k)}$), ou seja:

$$F_{hst(k)} = \frac{F_{a(k)} - \min(F_{a(k)})}{b_{w(k)}}, \tag{5.1}$$

e

$$F_{hsr(k)} = F_{a(k)} b_{w(k)} + \min(F_{a(k)}). \tag{5.2}$$

Em ambas as versões, F_{hst} e F_{hsr} representam, fundamentalmente, a informação inicial para realce e degradação. São utilizadas na obtenção das versões-guias iniciais para os *pipelines* do GFD e na geração da referência F_r. Embora o limiar de 1% seja satisfatório, F_a ainda pode ser afetado por pixels *outliers* em vídeos que contenham regiões onde há pixels que superem este limiar. Esta condição resulta em flutuações nos resultados para F_{hst} e F_{hsr} (histograma), produzindo *flickering*. Assim, o *frame* interno de referência F_r é determinado a partir da versão média entre F_{egf} e F_{dgf}, conforme indicado na Eq. 5.3. Atenua-se a influência dos *outliers* remanescentes sobre F_r através da fusão pela média. O procedimento via média aritmética determina resultado equidistante das versões F_{hst} e F_{hsr}, dado que ambas as operações, descritas nas Eqs. 5.1 e 5.2, são lineares.

$$F_{r(k)} = \frac{F_{hst(k)} + F_{hsr(k)}}{2}. \tag{5.3}$$

A obtenção dos *frames* efetivos de referência para o GFD, F_{egf} e F_{dgf}, é realizada através da média entre F_r e os *frames* F_{hst} e F_{hsr}, respectivamente, como indicam as Eqs. 5.4 e 5.5. Estas versões contém *features* duais e complementares associadas a qualidade da imagem e degradação. Em adição, mostram acentuada redução do efeito *flickering* entre transições de *frames*.

$$F_{egf(k)} = \frac{F_{r(k)} + F_{hst(k)}}{2}, \tag{5.4}$$

e

$$F_{dgf(k)} = \frac{F_{r(k)} + F_{hsr(k)}}{2}. \tag{5.5}$$

5.4.2 Módulo de Geração dos Frames-Duais

Este módulo tem estrutura similar à proposta para realce de imagem. É organizado em dois *pipelines* para a geração das versões-guias de realce e degradação, utilizadas na função de degradação propriamente dita. O *pipeline* para geração do *frame*-guia para realce conta com etapas de reforço sobre cor e contraste, seguido de etapa de reforço de textura. O *pipeline* para a geração do *frame*-guia para degradação compreende etapas de atenuação na cor e contraste, e para a redução de textura. As etapas de reforço e atenuação utilizam a representação no espaço de cor apresentada pela Eq. 4.5, na obtenção dos respectivos *frames*, conforme indicado nas Eqs. 5.6 e 5.7. $F_{ce(k)}$ corresponde à versão com reforço de cor e contraste de F_{egf} e F_{cd} é a versão com atenuação de F_{dgf}.

$$F_{ce(k)} = \frac{F_{egf(k)} - (1 - e^{-gd_{(k)}})\Lambda_{(k)}}{e^{-gd_{(k)}}}, \tag{5.6}$$

e

$$F_{cd(k)} = F_{dgf(k)}e^{-gd_{(k)}} + (1 - e^{-gd_{(k)}})\Lambda_{(k)}. \tag{5.7}$$

Os parâmetros gd e Λ são obtidos através das Eqs. 4.8 e 4.11 e, por conveniência, expressos para vídeo nas Eqs. 5.8 e 5.9.

$$gd_{(k)} = \sqrt{[\max(F_{r(k)}) - F_{r(k)}][F_{r(k)} - \min(F_{r(k)})]}, \tag{5.8}$$

e

$$\Lambda_{(k)} = \frac{\max(F_{r(k)}) + \min(F_{r(k)})}{2}. \tag{5.9}$$

A obtenção das versões-guias $Fe(k)$ e $F_{(d(k}$ é efetivada após as respectivas etapas de reforço e atenuação de textura, de forma idêntica à realizada na metodologia para imagem, sintetizadas nas Eqs. 5.10 e 5.11, onde $\mathcal{G}(\cdot)$ corresponde à filtragem gaussiana.

$$F_{e(k)} = 2F_{ce(k)} - \mathcal{G}(F_{ce(k)}), \tag{5.10}$$

e

$$F_{d(k)} = \frac{F_{cd(k)} + \mathcal{G}(F_{cd(k)})}{2}. \tag{5.11}$$

Finalizando, o *frame* $\hat{F}^*_{(k)}$ é obtido da função de degradação, ou seja:

$$\hat{F}^*_{(k)} = \hat{F}_{(k)} + F_{d(k)} - F_{e(k)} \tag{5.12}$$

6 METODOLOGIA DE IMPLEMENTAÇÃO

Neste capítulo são apresentados a metodologia associada à implementação, onde são discutidos o detalhamento da rede neural, a composição do *dataset* utilizado no trabalho, estratégia de treinamento e a função de perda. Importante destacar que estes aspectos de implementação são idênticos às metodologias para melhoria de imagem e vídeo, exceto na estratégia de treinamento para vídeo que foi estrutura em duas etapas.

6.1 Rede Neural

A bibliografia aponta as redes convolucionais como um padrão de fato para o processamento de imagens envolvendo DL. A escolha da topologia para a rede neural está vinculada ao arcabouço conceitual do método proposto. Por premissa, não são utilizados dados pareados para o treinamento e pré-processamento das imagens, fundamentando a opção pela abordagem por auto-aprendizado (SSL). O atendimento dos requisitos do método requer que o modelo DL extraia estruturas ocultas a partir dos dados não anotados, representando estas no espaço latente. Uma rede com topologia *encoder-decoder*, especificamente, um autoencoder realiza este procedimento enquanto tenta reconstruir sua entrada Pretorius et al. (2018). Objetivamente, redes autoencoder são estruturas simples de aprendizagem profunda, que procuram transformar a informação de entrada em saída com o mínimo de distorção, Baldi (2012). A escolha por esta arquitetura leva em conta estas características, associadas à concepção metodológica para melhoria de imagens UW proposta neste trabalho, onde se realiza uma estimativa da degradação da imagem de entrada, usando esta informação para conduzir o aprendizado da rede durante o treinamento.

A arquitetura do modelo implementado é mostrada na Fig. 6.1, onde cada bloco dimensional é indicado: número de camadas × largura × altura × número de mapas de *features*. A mesma rede é utilizada para o realce de imagem e vídeo, com as respectivas diferenças implementadas no Módulo de Degradação.

O modelo foi implementado utilizando os *frameworks* Keras/Tensorflow, sendo leve do ponto de vista computacional, com 170 mil parâmetros treináveis. Na seção *encoder* a redução de escala é realizada via convoluções com *stride* (2,2), evitando o uso de camadas de *pooling* e minimizando a perda de informação. Na seção *decoder* o aumento de escala é obtido via camadas de *upscaling* 2D com tamanho (2,2) e interpolação por vizinhos mais próximos (*nearest neighbors*). A opção por esta estratégia decorre dos resultados de Odena et al. (2016), que indicaram redução na geração de artefatos durante o aumento de escala. Os autoencoders apresentam limitações quando aplicados à restauração e ao realce de imagens, geradas principalmente na redução espacial da imagem nas camadas do *encoder*. Ronneberger et al. (2015) propôs o uso de atalhos conectando as seções *encoder-decoder*, identificados por *skip connections*. Estas estruturas foram implementadas na rede e fornecem informação tanto local como global no contexto da imagem para as camadas do *decoder*. Também, contribuem na redução

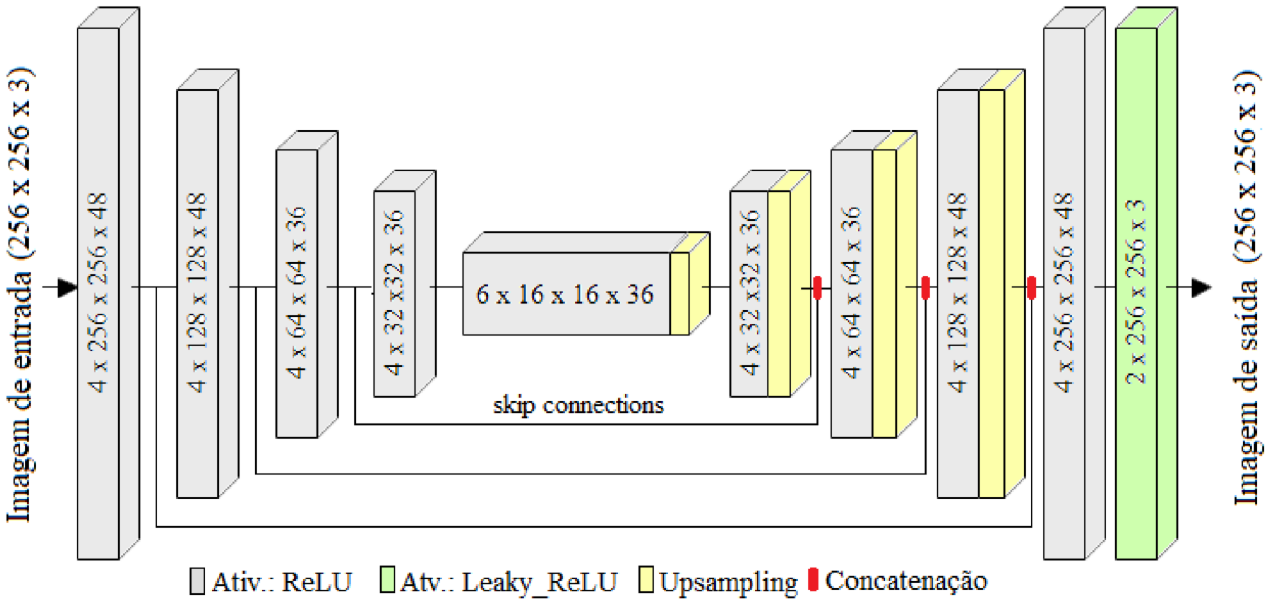


Figura 6.1: Arquitetura do autoencoder implementado

do risco de esvaecimento do gradiente (*vanishing gradients*) em redes profundas.

A ativação das camadas convolucionais é realizada por funções ReLU, dado serem computacionalmente simples, reduzindo o tempo de treinamento, com velocidade de convergência maior. Não saturam como as funções sigmoidal e tangente hiperbólica. Entretanto, nas camadas finais são utilizadas funções de ativação Leaky-ReLU, permitindo valores negativos para a ativação, evitando valores nulos para os gradientes quando a ativação resulta em valores menores que zero. Destaca-se que os hiperparâmetros da rede foram estabelecidos durante o desenvolvimento do modelo, como resultado de avaliações para melhoria do desempenho e soluções de problemas (ruído, tempo de treinamento, qualidade da reconstrução da imagem de saída, convergência da função de perda e de validação, etc.). Na Fig. 6.2 são mostrados um conjunto de imagens de entrada e as respectivas saídas, obtidas para o autoencoder treinado em 100 épocas, com mini-lote de 6 imagens. Na Fig. 6.3 é indicada a característica associado ao treinamento da rede.

As camadas de convolução usam regularização $L1$ com fatores de regularização de 15×10^{-6} para os pesos e 1.5×10^{-6} para o bias. Embora a regularização $L2$ seja normalmente a opção escolhida, dado que permite melhor acurácia, $L1$ tem melhor desempenho no ponto de vista da interpretabilidade. Uma propriedade da regularização $L1$ é o fato dela criar soluções esparsas, onde a maioria dos pesos é zero. Assim, se na camada de entrada uma dada conexão resulta em peso zero, então a respectiva entrada não tem efeito na predição. Ou seja, tal entrada pode ser descartada (*dropped*) e a regularização $L1$ atua como seletora de *features*, indicando quais são preditivas na aplicação, Charte et al. (2018), Aggarwal (2018). À primeira vista, esta propriedade aparenta ser bastante associada à tarefa de classificação. No entanto, Aggarwal (2018) indica um aspecto mais amplo, apontando que ela pode ser útil em aplicações onde o treinamento é realizado a partir de dados em que a natureza e as características gerais não se



Figura 6.2: Imagens de entrada e saída do autoencoder. (a) Entrada, (b) Saída.

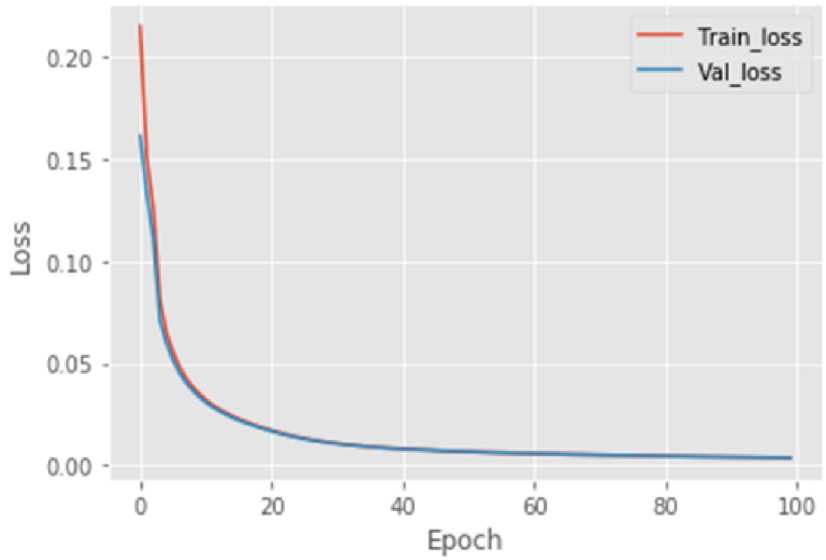


Figura 6.3: Característica de treinamento do autoencoder

alteram significativamente. Este aspecto é relevante no contexto do trabalho, pois considera-se que imagens (e vídeos) submersas mostrem similaridades em *features* associadas à turbidez e à interação com a luz refletida pelos objetos e com a luz ambiente (efeitos *forward* e *backscattering*, respectivamente). No trabalho, ambos os tipos de regularização foram testados. $L1$ mostrou melhores resultados em termos de percepção visual associada ao realce de vídeo. Entretanto, $L2$ resultou em maior convergência da função de perda para configurações de teste idênticas.

A inicialização é realizada utilizando-se o algorítimo *GlorotNormal* do Keras e o módulo de degradação foi implementado como uma camada não-treinável da rede.

6.2 *Dataset* UWData

As aplicações envolvendo modelos baseados em redes neurais dependem, fundamentalmente, do conjunto de dados que serão utilizados no treinamento. A efetividade nos resultados e a capacidade de generalização da rede são essenciais quando o objetivo é o processamento de imagens. A escolha ou adequação do conjunto de dados é essencial para o atendimento destas características.

A bibliografia disponibiliza vários *datasets* relacionados à imagens subaquáticas, que atendem diversas classes de aplicações como detecção, Liu et al. (2021) e Pedersen et al. (2019); classificação, Raine et al. (2020) e Boom et al. (2012); segmentação, semântica Yuval et al. (2021) e Li et al. (2021), mapeamento e localização, Zwilgmeyer et al. (2021) e Ferrera et al. (2019). Entretanto, a maioria destes *datasets* contém imagens com atributos específicos para a aplicação a que destinam, tais como: muito baixa turbidez, contexto monótono ou destacando apenas uma classe de objetos, ou contém imagens sintetizadas. No método proposto, a rede neural aprende a ação de realce durante o treinamento, utilizando a degradação como fonte de informação. Isto requer imagens naturais, apresentando condições ambientais (níveis de turbidez, *color cast*, luz ambiente) e cenários subaquáticos reais e diversificados. Importante destacar que o método não utiliza imagens pareadas, isto amplia e facilita a seleção de imagens.

No treinamento do modelo foram usadas imagens dos *datasets* UIEB, proposto por Li et al. (2020), SUIM proposto por Islam et al. (2020) e RUIE apresentado por Liu et al. (2020).

O *dataset Underwater Image Enhancement Benchmark (UIEB)* apresenta conjunto de 890 imagens naturais coletadas de publicações e repositórios na internet ou adquiridas pelos próprios autores. Embora não utilizado neste trabalho, este conjunto também contém as respectivas imagens de referência, obtidas a partir de testes subjetivos sobre os resultados de 12 métodos de realce apresentados na bibliografia, conforme descrito em Li et al. (2020). Na Fig. 6.4 são mostradas imagens exemplo do UIEB.



Figura 6.4: Amostras de imagens do *dataset* UIEB

Merece destaque a diversidade relacionada à turbidez e à luz ambiente nas imagens deste conjunto. No entanto, apresenta vários exemplos cuja aparência sugere edição, apresentando rótulos ou *tags* não pertencentes à cena e exemplos com exposição realizada em oposição a fonte de luz, particularmente, contra o sol, prejudicando bastante a percepção da cena.

O *dataset Segmentation of Underwater IMagery (SUIM)* contém 1500 imagens anotadas para treinamento de modelos para segmentação semântica. Em consequência, as imagens apre-

sentam menor degradação. Em geral, mostram melhores contraste e coloração que as propostas no conjunto UIEB. A Fig. 6.5 mostra exemplos das imagens contidas no SUIM.

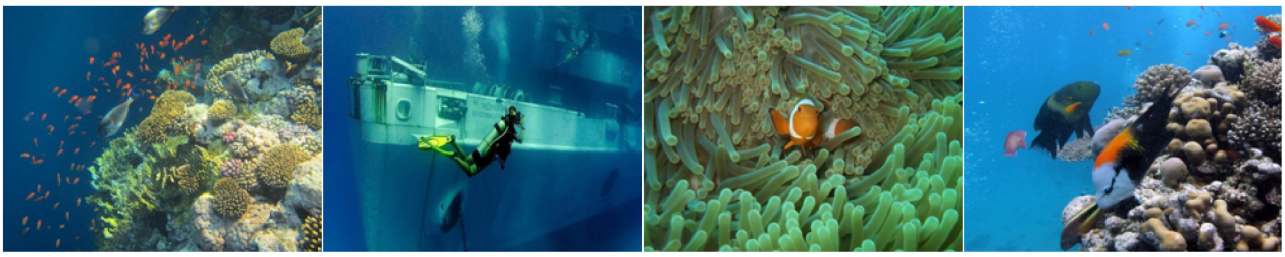


Figura 6.5: Amostras de imagens do *dataset* SUIM

O *dataset Real-world Underwater Image Enhancement (RUIE)* apresenta três subconjuntos, *Underwater Image Quality Set (UIQS), Underwater Colour Cast Set (UCCS)* e *Underwater Higher-level Task-driven Set (UHTS)*, destinados a avaliação da qualidade de imagem. Ao todo, contém 4230 itens. As imagens foram totalmente coletadas pelos autores, através de uma estrutura com 24 câmeras durante turnos de 3h em locais diferentes. Destacam-se as nuances geradas pela variação da maré e luz do sol. Entretanto, as condições ambientais, principalmente, no que tange ao *color cast*, são similares entre os subconjuntos. Na Fig. 6.6 são apresentados exemplos das imagens do RUIE.



Figura 6.6: Amostras de imagens do *dataset* RUIE

O *dataset* utilizado no treinamento dos modelos propostos para melhoria de imagem e vídeo, contém 2200 imagens. Foi construído utilizando-se 800 imagens do UIEB, 400 imagens do SUIM e 100 imagens do RUIE. Adicionalmente, foram coletadas 900 imagens e frames de vídeos disponíveis na internet. O conjunto final foi denominado de UWData[1]. Durante a seleção das amostras, priorizou-se imagens com níveis moderado e moderado-alto de turbidez e *color cast*. Estes fenômenos geram perda de contraste, atenuação de cor e borramento. Quando da interação com o luz ambiente, destacam o efeito *backscattering* com redução da percepção dos objetos na cena. Entretanto, imagens com níveis menores de degradação também foram selecionadas. Ao final, obteve-se aproximadamente, a composição de 30% de imagens com degradação baixa ou moderada-baixa, 50% de imagens com degradação moderada-alta e 20% com degradação alta.

Adicionalmente, foram utilizadas as imagens do *dataset* TURBID (Duarte et al., 2016) na etapa de avaliação quantitativa dos resultados do trabalho para imagem. O método não

[1] `https://drive.google.com/file/d/1jgzG9z68ESPsXYAUJeR5mobp3hJsIlNi/view?usp=drive_link`

requer imagens de referência ou *ground-truths* no treinamento do modelo, assim, a avaliação dos resultados fica dependente do aporte de métricas sem referência, ainda restrito dado o reduzido número opções para aplicações UW.

O TURBID é destinado à avaliação de métodos de melhoria de imagens UW. É composto por três subconjuntos de imagens subaquáticas, degradadas quimicamente de forma progressiva. Em adição, cada subconjunto contém imagem com a água totalmente limpa, sem turbidez, indicada como imagem de referência. Ao todo, o TURBID disponibiliza 66 imagens. Na Fig. 6.7 são amostradas imagens do subconjunto *Milk* do TURBID e a imagem de referência.



Figura 6.7: Amostras de imagens do *dataset*TURBID/Milk

6.3 Estratégias de treinamento

O treinamento do modelo para melhoria de imagens foi realizado em 300 épocas, com mini-lotes de 6 imagens. Os dados foram divididos randomicamente em 90% (80 + 10) para treino e validação e 10% para teste. Adicionalmente, utilizou-se o *dataset* UWTest[2] com 90 imagens não usadas no treinamento, destinadas à avaliação no estudo comparativo dos resultados obtidos com o modelo. Utilizou-se o algoritmo de otimização ADAM (Kingma and Ba (2017)), com taxa de aprendizagem configurada em 0.0008 e demais parâmetros nos seus valores *default* ($\beta_1 = 0,9$; $\beta_2 = 0,999$; $\epsilon_z = 1 \times 10^{-7}$).

O treinamento do modelo para melhoria de vídeo é realizado em duas etapas. A primeira é similar à executada para imagens, com o modelo para vídeo treinado com o UWData, em 300 épocas utilizando mini-lotes de 6 imagens e regularização $L1$. A segunda etapa é destinada ao ajuste do modelo ao vídeo-alvo (*fine tunning*). Durante a primeira fase do treinamento, o modelo aprende o realce a partir de uma diversidade de contextos, diferentes entre si. No realce de vídeo, o contexto tende a apresentar forte similaridade em termos ambientais. Nestas condições, o realce produzido vai depender apenas do que o modelo aprendeu com imagens de contextos similares na primeira fase. Isto pode ser bastante limitante, pois se o contexto do vídeo-alvo for contemplado em um número restrito de imagens no dados da primeira fase, o realce será limitado ao aprendizado obtido a partir destes dados.

Na execução da segunda fase do treinamento foram testadas e avaliadas três opções. A primeira consistiu na utilização integral de todos os *frames* do vídeo-alvo. Esta opção implica em alta probabilidade de sobre-ajuste (*overfitting*) do modelo aos dados. Entretanto, na

[2] `https://drive.google.com/file/d/11aVxTUlklT3CZeXjtD8PDli7RenZUGsk/view?usp=sharing`

perspectiva da aplicação, deseja-se obter o melhor desempenho possível de um sistema para melhoria da percepção visual de um vídeo UW e uma condição que conduza ao *overfitting* pode ser tolerada. As segunda e terceira opções para o treinamento usam apenas 20% dos *frames* do vídeo-alvo, explorando a baixa variabilidade de elementos de degradação em um vídeo subaquático contínuo, principalmente associadas à turbidez e ao *color cast*. Na segunda opção, foi utilizado trecho contínuo correspondendo a 20% do vídeo-alvo. Esta opção requer a seleção do trecho a ser utilizado no treinamento e, por princípio, qualquer segmento contínuo do vídeo poderia ser utilizado. Embora os resultados tenham sido idênticos à primeira opção, identificou-se diferenças de desempenho em determinados vídeos. A seleção do trecho de um trecho específico de um vídeo, reduz a capacidade de generalização do modelo para vídeos com intensas alterações tanto na dinâmica dos objetos nas cenas (vídeo "Octopus"), quanto na refletância destes objetos (vídeo "Mobula"). Para contornar esta dificuldade e minimizar a possibilidade de *overfitting*, a terceira opção testada para o treinamento utiliza 20% dos *frames* do vídeo contínuo, selecionados de forma descontínua e espaçados no tempo em intervalos regulares. Por exemplo, para uma sequência de vídeo de 1000 quadros, foram utilizados 200 *frames*, que correspondem à seleção de 1 em cada 5. O conjunto destes 200 *frames* foi empregado no *fine tunning*. Os resultados obtidos com esta opção foram idênticos aos da primeira opção nos vídeos testados. Todos os testes foram executados sob configuração para o treinamento da rede idêntica. O percentual de 20% escolhido resulta de testes, onde procurou-se definir o percentual mínimo necessário para garantir o ajuste do modelo para que o *fine tunning* permitisse resultados similares aos obtidos quando empregando todos os quadros do vídeo UW.

Os resultados apresentados neste trabalho correspondem ao modelo ajustado na segunda fase do treinamento conforme as condições da terceira opção. Assim, na segunda fase são utilizados, aproximadamente, 20% dos *frames* do vídeo-alvo para o ajuste via *fine tunning* do modelo. Os *frames* selecionados para esta etapa são não-contínuos. A efetivação do treinamento é então realizada em 10 épocas, com mini-lote de 6 *frames* e regularização $L1$. Os demais 80% dos *frames* não utilizados no treinamento, são utilizados como dados de teste. Este conjunto de 20% são suficiente para o ajuste do modelo ao vídeo-alvo, sendo executado com baixo consumo de tempo, o que é tema importante para a aplicação. Um ROV, por exemplo, pode ter embarcado o modelo pré-treinado, via primeira etapa, e realizar a segunda *in situ*, imediatamente antes do início da missão.

6.4 Função de perda

A função de perda é baseada no Erro Quadrático Médio (*Mean Square Error - MSE*). Consiste em uma função de perda bastante popular, principalmente em abordagens envolvendo problemas de regressão e imagem. A expressão geral é indicada na Eq. 6.1. Apresenta bom desempenho quando associada à imagens com pequenas dimensões e a algorítimos com processamento baseado em pixels, Khare et al. (2022).

$$\mathcal{L} = \frac{1}{N}\sum_{t=1}^{N} ||I_{ref} - I_{pred}||_2^2, \tag{6.1}$$

onde N é o número de pixels da imagem, I_{ref} é a imagem de referência e I_{pred} e a saída da rede. No trabalho, I_{ref} corresponde à imagem de entrada $I_{(k)}$ e I_{pred} corresponde à imagem resultante do módulo de degradação $\hat{I}^*_{(k)}$.

Na implementação, a função de perda foi concebida em dois termos. Um termo associado a radiância da imagem como um todo, e outro relacionado à radiância do contexto dos objetos na cena. O primeiro fornece informação global durante a decodificação (reconstrução), enquanto o segundo tem o objetivo de destacar os objetos presentes na cena.

O termo da refletância dos objetos na cena é determinado a partir da representação proposta na Eq. 4.5 e interpretada a partir de uma soma vetorial. Ou seja, os termos $I_{J(k)}(x,y)e^{-gd_{(k)}(x,y)}$ e $(1-e^{-gd_{(k)}(x,y)})\Lambda_{(k)}$ na Eq. 4.5 são vetores, que compõem uma ação complementar ponderada pelo termo exponencial, como expresso na Eq. 6.2. Neste contexto, define-se a imagem como o vetor resultantes destas duas componentes.

$$\overrightarrow{I}_{(k)} = \overrightarrow{I}_{J(k)}e^{-gd_{(k)}} + (1-e^{-gb_{(k)}})\overrightarrow{\Lambda}_{(k)}. \tag{6.2}$$

O diagrama vetorial desta operação e ilustrado na Fig. 6.8. Assim, se redefinindo a Eq. 6.2,

$$I_{(k)} = I_{J(k)}cos(\alpha_{(k)}) + \Lambda_{(k)}cos(\beta_{(k)}), \tag{6.3}$$

onde $I_{(k)} = |\overrightarrow{I}_{(k)}|$, $I_{J(k)} = |\overrightarrow{I}_{J(k)}|$, $\Lambda_{(k)} = |\overrightarrow{\Lambda}_{(k)}|$ e:

$$cos(\alpha_{(k)}) = e^{-gd_{(k)}}, \tag{6.4}$$

e

$$cos(\beta_{(k)}) = (1-e^{-gd_{(k)}}). \tag{6.5}$$

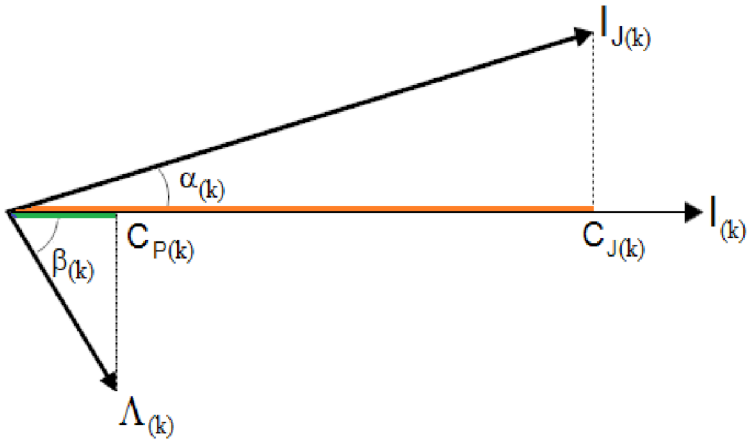


Figura 6.8: Diagrama da representação da imagem interpretada como uma composição vetorial

Na Figura 6.8, C_{Jk} e C_{Pk} são as componentes de $I_{J(k)}$ e $\Lambda_{(k)}$ na direção de $I_{(k)}$, respectivamente, e podem ser expressas por:

$$C_{J(k)} = I_{J(k)}.cos(\alpha_{(k)}), \tag{6.6}$$

e

$$C_{P(k)} = \Lambda_{(k)}.cos(\beta_{(k)}). \tag{6.7}$$

Assim, $I_{(k)}$ pode ser definida como:

$$I_{(k)} = C_{J(k)} + C_{P(k)}. \tag{6.8}$$

A abordagem acima pode ser estendida para a imagem de saída do módulo de degradação, $\hat{I}^*_{(k)}$. Obtendo-se:

$$\hat{I}^*_{(k)} = \hat{C}^*_{J(k)} + \hat{C}^*_{P(k)}. \tag{6.9}$$

Nas Eqs. 6.8 e 6.9, as componentes $C_{P(k)}$ e $\hat{C}^*_{P(k)}$ representam frações da luz ambiente em $I_{(k)}$ e $\hat{I}^*_{(k)}$ que incidem sobre a lente da câmera. E as componentes $C_{J(k)}$ e $\hat{C}^*_{J(k)}$ representam a informação relativa aos objetos na cena sem a influência do *forward scattering*. As imagens $I_{(k)}$ e $\hat{I}^*_{(k)}$ são comparadas na função de perda durante o treinamento. Assim, ao invés de usar apenas um termo definido em função destas imagens, é proposto termo adicional na função de perda, composto pelas componentes $C_{J(k)}$ e $\hat{C}^*_{J(k)}$, obtidas como indicado em (6.10) e (6.11), respectivamente.

$$C_{J(k)}(x, y) = I_{(k)}(x, y) - C_{P(k)}(x, y), \tag{6.10}$$

e

$$\hat{C}^*_{J(k)}(x, y) = \hat{I}^*_{(k)}(x, y) - \hat{C}^*_{P(k)}(x, y). \tag{6.11}$$

A componentes $C_{P(k)}$ é obtida de 6.7 e $\hat{C}^*_{P(k)}$ é obtida da expressão análoga, mas descrita a partir de seus parâmetros. Concluindo, os termos da função de perda são expressos conforme as Eqs. 6.4 e 6.4.

$$\mathcal{L}_i(I_{(k)}, \hat{I}^*_{(k)}) = \frac{1}{N}\sum_{t=1}^{N} ||I_{(k)} - \hat{I}^*_{(k)}||_2^2,$$

$$\mathcal{L}_i(I_{(k)}, \hat{I}^*_{(k)}) = MSE(I_{(k)}, \hat{I}^*_{(k)}). \tag{6.12}$$

E

$$\mathcal{L}_c(C_{J(k)}, \hat{C}^*_{J(k)}) = \frac{1}{N}\sum_{t=1}^{N} ||C_{J(k)} - \hat{C}^*_{J(k)}||_2^2,$$

$$\mathcal{L}_c(C_{J(k)}, \hat{C}^*_{J(k)}) = MSE(C_{J(k)}, \hat{C}^*_{J(k)}). \tag{6.13}$$

A expressão final para a função de perda é indicada na Eq. 6.14. O termo associado à

imagem fornece informações de contexto, como luz de fundo e contraste em relação aos objetos. O termo associado à cena favorece o destaque de cor e contorno dos objetos.

$$\mathcal{L} = c_1 \mathcal{L}_i(I_{(k)}, \hat{I}^*_{(k)}) + c_2 \mathcal{L}_c(C_{J(k)}, \hat{C}^*_{J(k)}). \tag{6.14}$$

Os fatores de ponderação c_1 e c_2 foram ajustados em $c_1 = 0,4$ e $c_2 = 0,6$, após avaliação experimental.