

A Semantic-oriented Approach for Underwater Wireless Communications using Generative AI

João Pedro Loureiro, Afonso Mateus, Filipe B. Teixeira, Rui Campos
INESC TEC and Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, s/n - 4200-465 Porto, Portugal
{joao.p.loureiro, afonso.mateus, filipe.b.teixeira, rui.l.campos}@inesctec.pt

Abstract—Underwater wireless communications are crucial for supporting multiple maritime activities, such as environmental monitoring and offshore wind farms. However, the challenging underwater environment continues to pose obstacles to the development of long-range, broadband underwater wireless communication systems. State of the art solutions are limited to long range, narrowband acoustics and short range, broadband radio or optical communications. This precludes real-time wireless transmission of imagery over long distances.

In this paper, we propose SAGE, a semantic-oriented underwater communications approach to enable real-time wireless imagery transmission over noisy and narrowband channels. SAGE extracts semantically relevant information from images at the sender located underwater and generates a text description that is transmitted to the receiver at the surface, which in turn generates an image from the received text description. SAGE is evaluated using BLIP for image-to-text and Stable Diffusion for text-to-image, showing promising image similarity between the original and the generated images, and a significant reduction in latency up to a hundred-fold, encouraging further research in this area.

I. INTRODUCTION

Sea-based activities are constantly growing, with offshore wind farms, environmental monitoring, and deep-sea mining being the most prominent examples [1]. However, the challenging nature of the ocean makes the development of long-range, broadband underwater wireless communications systems an unsolved problem. Due to the propagation characteristics of the water, underwater wireless communications are limited either to broadband, short-range solutions or narrowband, long-range solutions [2]. Two approaches are used for short-range scenarios: 1) radio-based communications, which can achieve high throughput but the water conductivity causes a strong attenuation of the electromagnetic waves, limiting its range; and 2) optical communications, which can also achieve high throughput, but are easily affected by water turbidity. For long-range scenarios, acoustic communications are used, allowing the transmission of information at tens of kilometers, but at a very low throughput. This prevents the real-time wireless transmission of data, such as images, over long distances captured by robotic platforms such as Autonomous Underwater Vehicles (AUVs).

A new paradigm is emerging in broadband wireless communications, which consists of the use of semantics to transmit

data. The key idea is to extract semantic features of the information to be sent, and then forward those features to the receiver instead of the original information (e.g. the image), reducing the size of the data sent. Semantic communications appear as a promising approach for underwater communications, potentially enabling a more efficient use of the limited bandwidth available, especially for acoustic links. Additionally, the use of a semantic-oriented approach can improve error tolerance, as it can be possible for the receiver to understand the semantics of the received information even if some data is lost or corrupted. This is especially relevant for intermittent and unreliable wireless underwater channels, such as acoustics [3].

The main contribution of this paper is SAGE, an innovative semantic-oriented approach for real-time underwater wireless imagery transmission over narrowband channels. SAGE combines Generative Artificial Intelligence (genAI) capabilities of the sender node underwater to extract semantically pertinent information from images, which are transmitted in text format. This highly compressed information is sent to the receiver, which also uses genAI capabilities to generate an image from the received text description. To evaluate SAGE performance, a testbed was created based on BLIP [4] and Stable Diffusion [5] algorithms to test image-to-text and text-to-image conversion, respectively. Laboratory tests show promising image similarity results and reduced communication latency.

The rest of the paper is organized as follows. Section II reviews the state of the art in underwater communications and semantic algorithms. Section III presents SAGE and explains its rationale. Section IV describes the results for delay performance and image similarity. Section V draws the conclusions and points out the future work.

II. STATE OF THE ART

This section provides an overview on existing underwater wireless communication technologies and recent advances in underwater semantic communications. We also present some semantic image-to-text and text-to-image solutions.

A. Underwater Wireless Communications

Underwater wireless communications mainly rely on three technologies: acoustic, optical, and radio [2]. Acoustic solutions use sound waves to exchange messages underwater at long distances, but they are limited to low data rates. Optical communications, while cost-effective and supporting high data rates at short ranges, are vulnerable to interference

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020 and the scholarship 2022.14283.BD.

from water turbidity and generally require line-of-sight. Radio communications can achieve high data rates but suffer from strong signal attenuation, which results in ranges in the order of tens of centimeters.

Alternative approaches have been proposed in literature to overcome the limitations of these three technologies, such as data muling and multimodal solutions. Data muling was explored in several works such as in [6], consisting in the use of one or more AUVs as data mules, that physically transport the data between two distant nodes. This enables the delivery of large amounts of data using high data rate, short-range technologies (e.g. radio or optical), when the mule is close to the target node. Despite this, the time the mule takes to travel between the two physical points and the need of localization and navigation mechanisms is a challenge. Another alternative is a multimodal approach, integrating different technologies into a single system, minimizing the limitations of each technology and maximizing its main potentialities [7]–[9]. In [8], the DURIUS system tries to combine radio, optical and acoustic technologies, with the goal of reducing both overall delay and energy consumption. Nonetheless, it does not address the real-time wireless transmission of imagery over long distances.

B. Underwater Semantic Communications

Semantic communications refer to the transmission of meaning of the information rather than the raw data, in order to improve the efficiency and effectiveness of communication systems. In traditional communication systems, information is sent bit by bit, whereas in semantic communications, only the essential meaning or semantic content is transmitted. As a result, this approach significantly reduces the required communication resources, namely bandwidth [10]. Semantic communication relies heavily on AI, making it a communication model predominantly dependent on AI and the corresponding high-processing requirements.

The use of a semantic approach to improve wireless communications underwater is still very limited. To the best of our knowledge, only two works have been published so far. In [11], the authors propose using semantic communications to improve the reliability and efficiency of underwater acoustic links. A simple simulation model is proposed, as well as suggestions for future work, pointing out the need to adapt models to changeable underwater conditions. Likewise, a deep-learning-based underwater wireless optical semantic communications system is proposed in [12]. The system improves image transmission efficiency by extracting and transmitting only basic semantic information, with experimental validation demonstrating superior performance when compared to traditional methods. These works have only considered initial designs and testing, and the challenge of implementing a semantic-oriented solution for long-range scenarios was not yet explored.

C. Semantic Image-to-Text Solutions

A number of language models offer the capability to generate textual descriptions based on semantic information. CLIP [13] employs a contrastive learning approach, which aligns images with their corresponding text descriptions by maximizing the similarity between correct pairs and minimizing the

similarity for incorrect pairs [13]. This approach enables CLIP to perform in zero-shot learning, i.e., recognize unseen objects by using descriptive attributes instead of examples, demonstrating high versatility across various applications. The ViLBERT model [14] combines image and text features using two parallel Bidirectional Encoder Representations from Transformers (BERT) streams, where one stream processes the image and the other processes the text. These streams interact through Co-Attentional Transformer Layers, enabling the model to learn potential relationships between visual and linguistic elements. ViLBERT is particularly effective in tasks requiring deep understanding of the interactions between objects in an image and their descriptions, such as Visual Question Answering and image captioning. Finally, BLIP [4] employs a dual-stream approach, utilizing separate transformers for processing images and text independently. These separate streams are unified through a series of Co-Attentional Transformer Layers, similar to ViLBERT, allowing the model to effectively merge visual and textual information. BLIP is particularly noteworthy for its bootstrapping technique that allows the learning from noisy image-text pairs, i.e. those containing mismatched errors, thereby demonstrating exceptional robustness and adaptability even when handling imperfect data.

D. Semantic Text-to-Image Solutions

Recent advancements in the text-to-image domain have led to the development of sophisticated models that generate images from text descriptions. DALL-E [15] employs a variant of the transformer architecture specifically tailored for generating images from text prompts. By training on a massive dataset of text-image pairs, DALL-E learns to map textual descriptions to visual representations through a combination of autoregressive and transformer models. This allows DALL-E to generate highly detailed and diverse images from a wide range of textual inputs. Imagen [16] leverages a combination of large-scale transformer models and a cascade of super-resolution networks to generate images from text. The model first generates a low-resolution image based on the textual description and subsequently refines it through a series of super-resolution steps to enhance quality and detail. Imagen excels in applications requiring precise and detailed visual outputs, benefiting from extensive pre-training on diverse and high-quality datasets. Finally, Stable Diffusion [5] is a generative model that employs the latent diffusion model architecture, operating in a compressed latent space. This model is based on the concept of reversing the progressive addition of noise to images during its training phase. The training process of Stable Diffusion involved a large dataset, enabling the model to learn the details required to generate images by effectively reversing the noise addition process. As it operates within a lower-dimensional latent space, Stable Diffusion achieves high computational efficiency while ensuring the generation of high-quality images. Furthermore, as an open-source model, it allows for community contributions and adaptations, making it suited for both research and practical applications.

Semantic-oriented communications are a promising approach to overcome the limitations of current underwater communications systems. Yet, current research targeting underwater scenarios is still in its infancy. Concerning semantic algorithms, as we have seen in this section, great steps have

been taken in the fields of both image-to-text and text-to-image algorithms, but none of them is specifically tailored for underwater context. Therefore, a new approach is still required to implement a reliable underwater semantic-oriented communications system.

III. PROPOSED SEMANTIC-ORIENTED APPROACH

The proposed semantic-oriented approach – SAGE – is focused on a point-to-point underwater wireless link, as shown in Fig. 1. The sender node (e.g. an AUV) is typically underwater, collecting data images of the environment, while the receiver node is generally close to surface (e.g. a vessel or an Autonomous Surface Vehicle). The wireless link can utilize acoustic, radio or optical technologies, with acoustics being the primary focus due to its longer-range capabilities.

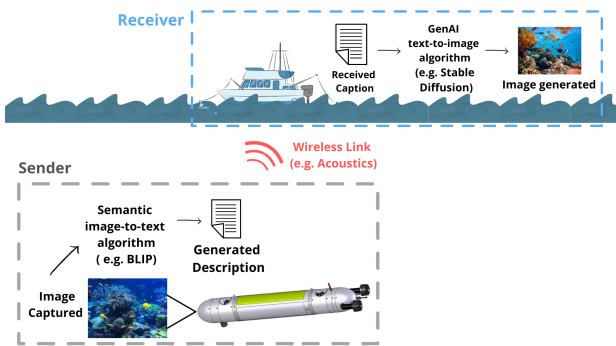


Fig. 1: SAGE high-level architecture, from image capturing at the sender, to image generation at the receiver.

The instance running on the sender node processes the images being captured by the AUV, generating a text description using an image-to-text algorithm. SAGE currently uses BLIP for its relative simplicity and ability to produce concise descriptions. These are relevant aspects concerning the limited computing power available in the AUV and the narrowband wireless channels, such as acoustics. At the other end, the receiver node picks up the text description transmitted by the sender via the wireless link and generates an image accordingly. Although it is a resource-consuming solution, SAGE currently uses Stable Diffusion to generate the images at the receiver since it is one of the most reliable text-to-image algorithms available. SAGE can use a remote node (e.g. cloud servers) to overcome the processing demand of this algorithm, assuming the receiver, at surface, is connected to the Internet.

In order to evaluate SAGE using BLIP as the image-to-text algorithm and Stable Diffusion as the text-to-image algorithm, we implemented a two-node testbed, using two machines. The sender node was implemented in a PC with a regular dual-core CPU, without GPUs attached, while the receiver node ran in a computer with a T4-GPU. The communication link was implemented with an Ethernet connection between the two nodes. We developed two applications: 1) one at the sender node, which generates a description for each image using BLIP and then forwards those descriptions to the receiver; and 2) one at the receiver node, that uses the received descriptions to generate images with Stable Diffusion.

IV. RESULTS AND DISCUSSION

To evaluate SAGE, we have first studied the implications of this approach on delay when compared with traditional image transmission. Then we have made an analysis of image similarity when using SAGE.

A. Implications on Communications Performance

Delay is an important metric in underwater systems, especially when very low bitrate channels are used. The reduction of delay achieved when using SAGE for imagery transmission was evaluated considering an ideal wireless link between the sender and the receiver. The delay for sending data (Δt_{trans}) over the wireless link is given by Eq. 1.

$$\Delta t_{trans} = \frac{Data_Size}{Bitrate} \quad (1)$$

In Eq. 1, $Data_Size$ is the size of the data (e.g. image) to be sent in bits, and $Bitrate$ is the bitrate associated with the wireless channel. Following the same reasoning, it is possible to calculate the time required for sending the information using SAGE, defined by Eq. 2. In this case, the data to be sent is the text description generated by BLIP. To the time required to send the text description we need to add the time for: 1) generating the description, at the sender (Δt_{gen_desc}); and 2) generating the image, at the receiver (Δt_{gen_image}).

$$\Delta t_{SAGE} = \Delta t_{gen_desc} + \Delta t_{trans} + \Delta t_{gen_image} \quad (2)$$

We plot the delay results in Fig. 2, with 95% confidence intervals, considering a range of bitrates consistent with the acoustic modems currently available [3]. We considered an image size of 500 kByte, to compare the performance of a traditional acoustic communication and SAGE. Using the two-node testbed referred in Sec. III, we obtained the average time to generate a text description (≈ 10 s), the average time to generate an image (≈ 60 s), and the average size of a text description (≈ 70 Byte).

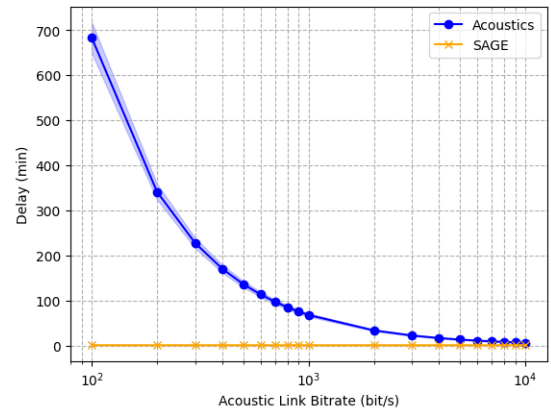


Fig. 2: Delay comparison for a 500 kByte image transmission over an acoustic link and SAGE.

As we can see, in terms of delay SAGE can achieve a delay reduction factor of $\frac{400s}{70.056s} \approx 5.7$ when considering a 10 kbit/s

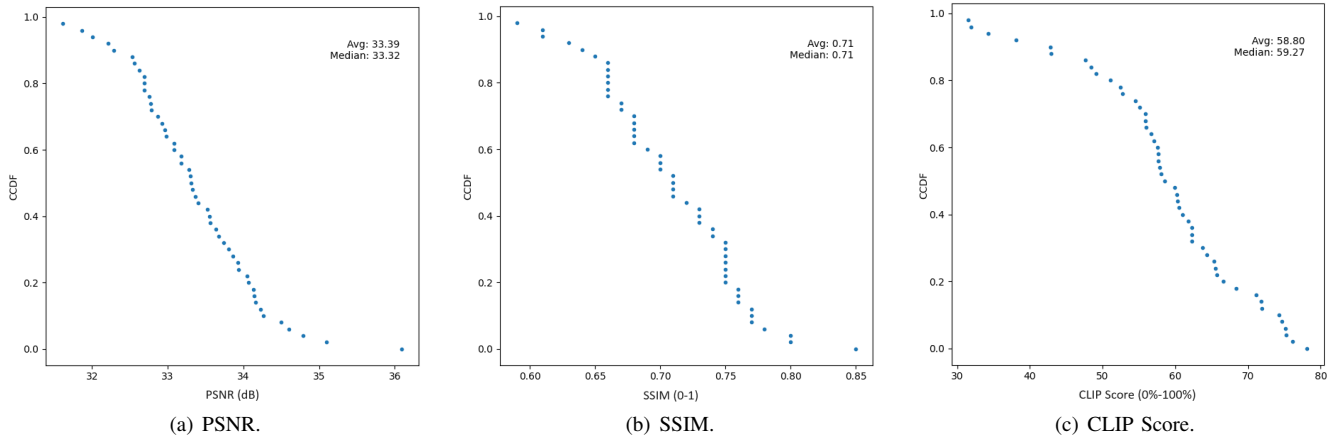


Fig. 3: CCDFs for Image Similarity Metrics.

bitrate, $\frac{800s}{70.112s} \approx 11.41$ at 5 kbit/s bitrate, and $\frac{4000s}{70.56s} \approx 56.70$ at 1 kbit/s bitrate. Thus, the latency reduction is particularly significant for bitrates below 1 kbit/s, which are typical in long-distance acoustic links. Even though the information being delivered by semantic-oriented communications is not visual data, the text description sent contains the basic semantic information, which enables the receiver to generate similar images. Despite the many advantages over traditional acoustic-based communication, it is important to note that the delays observed with SAGE are not yet compatible with real-time. To achieve true real-time performance, we need to reduce the image-to-text and text-to-image conversion times, as these are the main factors that contribute to latency when using the SAGE approach. This will be the subject of the future work.

B. Image Similarity Analysis

Different metrics have been proposed to measure the similarity between images, particularly for evaluating the efficiency of compression algorithms. Structural Similarity Index (SSIM) [17] and Peak Signal-to-Noise Ratio (PSNR) [18] are established metrics widely employed for assessing image quality and similarity. SSIM quantifies the perceptual difference between two images by evaluating alterations in structural information, luminance, and contrast, aiming to approximate human visual perception. In contrast, PSNR quantifies the ratio between the maximum possible power of a signal and the power of corrupting noise, expressed in dB, and is based on the Mean Squared Error (MSE) between the reference and test images. Although these metrics have been extensively used for comparing image quality, both are limited when applied to images generated using semantic algorithms. SSIM and PSNR predominantly focus on pixel-level discrepancies and low-level features, potentially failing to accurately capture the semantic content and perceptual quality of these images. For instance, an image generated by a semantic algorithm may exhibit substantial structural deviations from a reference image while maintaining perceptual similarity. Consequently, SSIM and PSNR may be inadequate for analysing the true visual and semantic similarity between images. This led to the need of creating alternative metrics that can effectively assess the perceptual and semantic aspects of images. CLIP Score [19]

offers a significant advantage over SSIM and PSNR, when comparing semantic content. As explained in Sec. II, CLIP leverages a model trained on image-text pairs to understand the meaning and the context of images. Besides text-image accuracy evaluation, CLIP Score can be adapted to compare two images, starting by computing the embeddings of the two images and then calculating the so-called cosine similarity between the embeddings [20].

To analyze the similarity between generated images at the receiver and the original images at the sender node, we present in Fig. 3 the Complementary Cumulative Distribution Function (CCDF) for SSIM, PSNR, and CLIP Score. CCDF was used because this type of function is the most effective way to analyze the distribution of similarity scores. Our set-up of tests consisted of 50 selected images from the dataset available at [21]. For each selected image, we repeated the process ten times, in order to reduce the random effects of Stable Diffusion at the receiver, since it generates different images from the same description.

PSNR values typically range from 20 to 80 dB, where higher values indicate higher similarity. As we can see in Fig. 3(a), the average PSNR reaches only 33.39 dB. SSIM values vary from 0 to 1, with 1 representing a perfect match. If we look to Fig. 3(b), we got an average and median of 0.71. For the CLIP Score (Fig. 3(c)), the average and the median are approximately 59%.

When we compare the similarity between the original and the generated images, we can conclude that these algorithms are still not very effective when the image refers to a specific underwater scenario. Nonetheless, the results for SSIM and CLIP Score show that, on average, the images are reasonably similar, accurately generating identical semantic content. On the other hand, we can observe that the metrics used are inconsistent between them. Although SSIM is a slightly more sophisticated metric than PSNR, it still relies on the structural appearance of the image, which may not be the best approach to evaluate the semantic similarity of the images. CLIP Score seems to be a more suitable metric for this purpose, since it better fits human perception of the image content.

Fig. 4 presents one of the sample pairs considered to plot



Fig. 4: Comparison of a sample image (original vs. generated).

TABLE I: Similarity metrics for the image in Fig. 4

Generated Caption (BLIP)	PSNR	SSIM	CLIP Score
<i>A photography of a wreck in the ocean</i>	34.22dB	0.72	64.14%

the CCDFs. For this pair of images, we can say that, despite the difference in the colors and in the background, the semantic information – *a ship wreck in the ocean* – can be observed in both the original and the generated images. The description length is variable and it is defined by BLIP. Table I presents the values for the three selected metrics. SSIM and CLIP Score values are reasonably close and roughly match the human perception of similarity between the two images. On the other hand, the PSNR values indicate low similarity, which is true if we consider the structural information and not the semantics of the visual data.

These results show that SAGE is a promising alternative for delay reduction in image transmission, enabling the receiver to obtain visual information in a drastically shorter time. This is a crucial aspect in underwater environments, where currently there is no solution for wireless image transmission in real-time, namely at long distances. In addition, the obtained results suggest that the selection of similarity metrics should be carefully considered according to the similarity analysis taking place and the operating scenario. In our particular case, PSNR appears to be unsuitable for differentiating semantic nuances, while CLIP Score seems to be the most accurate when compared to the human perception.

V. CONCLUSIONS AND FUTURE WORK

We presented SAGE, a semantic-oriented underwater communications approach to enable real-time wireless imagery transmission over noisy and narrowband channel, such as underwater acoustic networks. From the obtained results, we can conclude that SAGE can significantly impact the efficiency of an underwater wireless link, demonstrating a potential for significant latency reduction compared to traditional communication methods. By delivering semantic information instead of the original image data, SAGE optimizes underwater communications performance and paves the way for next-generation underwater wireless communications. Our findings suggest that BLIP and Stable Diffusion require adaptation for optimal performance in underwater environments. Furthermore, they suggest that image similarity metrics should be carefully selected, namely considering the operating scenario.

As future work, we plan to consider other image-to-text and text-to-image algorithms at the sender and receiver, and fine-tune genAI algorithms (BLIP, Stable Diffusion, or others) with a large set of underwater images, in order to enhance SAGE's efficiency and reliability.

REFERENCES

- [1] D. B. Alves, *Blue Economy*. Fundação Francisco Manuel dos Santos, January 2022, ISBN: 9789899004955.
- [2] S. Li, W. Qu, C. Liu, T. Qiu, and Z. Zhao, "Survey on high reliability wireless communication for underwater sensor networks," in *Journal of Network and Computer Applications*, 2019.
- [3] S. Sendra, J. Lloret, J. M. Jimenez, and L. Parra, "Underwater acoustic modems," *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4063–4071, 2016.
- [4] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," *CoRR*, 2022.
- [5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CoRR*, 2021.
- [6] J. P. Loureiro, F. B. Teixeira, and R. Campos, "Adaptive and Reliable Underwater Wireless Video Streaming using Data Muling," in *IEEE/MTS OCEANS 2021, Porto and San Diego*, 2021.
- [7] J. Liu, J. Wang, S. Song, J. Cui, X. Wang, and B. Li, "MMNET: A multimodal network architecture for underwater networking," *Electronics*, vol. 9, no. 12, 2020.
- [8] J. P. Loureiro, F. B. Teixeira, and R. Campos, "DURIUS: A Multimodal Underwater Communications Approach for Higher Performance and Lower Energy Consumption," in *2023 IEEE 9th WF-IoT*, 2023.
- [9] G. Moreira, J. P. Loureiro, F. B. Teixeira, and R. Campos, "Aquacom: A multimodal underwater wireless communications manager for enhanced performance," in *2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON)*, 2024, pp. 914–919.
- [10] Z. Qin, X. Tao, J. Lu, and G. Y. Li, "Semantic communications: Principles and challenges," *CoRR*, 2022.
- [11] J. Zhang, W. Sun, Y. Zhao, and H. Du, "Semantic communication in underwater communication: Advantage, problem and solution — a survey," in *2023 8th ICSP*, 2023, pp. 2120–2123.
- [12] J. Xu, Z. Huang, Y. Gao, W. Zhai, H. Qiu, and Y. Ji, "Design and experimental demonstration of underwater wireless optical communication system based on semantic communication paradigm," *Opt. Express*, vol. 32, no. 2, pp. 2188–2201, Jan 2024.
- [13] A. R. et al., "Learning transferable visual models from natural language supervision," *CoRR*, vol. abs/2103.00020, 2021.
- [14] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *CoRR*, vol. abs/1908.02265, 2019.
- [15] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *CoRR*, vol. abs/2102.12092, 2021.
- [16] C. S. et al., "Photorealistic text-to-image diffusion models with deep language understanding," 2022. [Online]. Available: <https://arxiv.org/abs/2205.11487>
- [17] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [18] D. R. Bull and F. Zhang, "Chapter 4 - digital picture formats and representations," in *Intelligent Image and Video Compression (Second Edition)*. Oxford: Academic Press, 2021.
- [19] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIP-Score: A reference-free evaluation metric for image captioning," in *2021 EMNLP*, Nov. 2021, pp. 7514–7528.
- [20] "Unlocking OpenAI CLIP. Part 2: image similarity," [Accessed: 14/08/2024]. [Online]. Available: <https://medium.com/@jeremyk/unlocking-openai-clip-part-2-image-similarity-bf0224ab5bb0>
- [21] L. Peng, C. Zhu, and L. Bian, "U-shape transformer for underwater image enhancement," *IEEE Transactions on Image Processing*, vol. 32, pp. 3066–3079, 2023.