Construction of a cultural image visual perception and aesthetic model based on a deep learning algorithm

Jinsong He and Syamsul Bahrin Zaibon
School of Creative Industry Management and Performing Arts, Universiti Utara Malaysia, Malaysia

ABSTRACT

This article studies the perceptual quality evaluation method of humanistic images based on the characteristics of human vision. First, this study obtained the disparity map of the human visual cultural image and used the edge map and saliency map to correct the weight of the disparity map. The weighted disparity map was then integrated into a perceptible score using a Minkowski fused deep neural network. A multi-scale method was employed to process the image to obtain the visual effect score. Taking Polytechnic University of Lausanne (EPFL) as an example, an experimental study was conducted on the three-dimensional visual quality assessment method proposed in this article. The test results with the EPFL library show that the target score obtained by this method has good consistency and monotonicity with the subjective score of the EPFL library. The article's results verify the effectiveness of the visual effects and aesthetics evaluation method proposed in this article.

KEYWORDS Stereoscopic image; deep learning algorithm; aesthetic model; visual perception; edge detection; multi-scale
CONTACT Jinsong He HJS20210302@126.com
Received 19 March 2024

1. Introduction

In recent years, with the rise of many industries, such as entertainment and the military, 3D imaging has been accepted by more and more people. Compared with 2D images, 3D images can bring more depth of field to viewers and provide them with an immersive visual experience. This helps people have a more detailed and accurate understanding of the content expressed in the image (Xin et al., 2022). How to evaluate it quickly, accurately and efficiently is an urgent problem that needs to be solved. However, factors unique to 3D imaging, such as crosstalk, adjustment-vergence conflict, step deformation, inconsistent depth of field, ghost effect, cardboard effect, etc., will also adversely impact the effect of 3D imaging. Therefore, there is a big difference in the quality evaluation of 3D and 2D images (Bonnen et al., 2021). Current research in this field is still in the process of exploration. There is no consensus among academic circles on the various factors involved in eye comfort. However, in the existing research results, the mechanisms of regulation–convergence conflict, crosstalk, staircase distortion, ghost effect, etc., are still unclear (Tsuneki, 2022). Existing research primarily uses eye-tracking instruments to evaluate human visual fatigue, which is challenging to detect effectively under non-human intervention conditions (Funke et al., 2021). This paper presents a detailed study of the three-dimensional imaging quality evaluation data in the EPFL system. A reference-free stereoscopic visual effect evaluation method was designed based on the characteristics of the human eye. Among them, technologies such as visual difference, edge extraction, saliency detection and multi-scale decomposition play a crucial role. Finally, the EPFL three-dimensional visual quality evaluation library was used to test the proposed algorithm to prove the rationality and practicality of the method proposed in this project.

2. Method

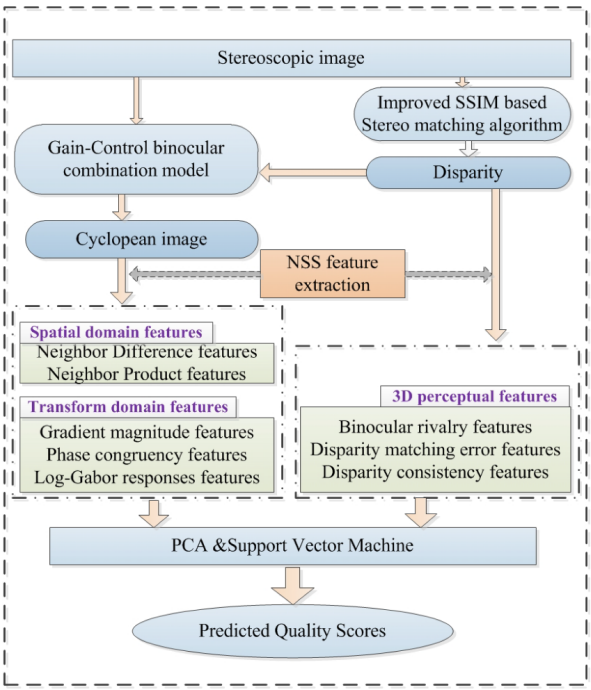

Figure 1. Perceptual quality evaluation algorithm.

Compared with 2D imaging, 3D imaging pays more attention to the effect of stereoscopic imaging rather than simply focusing on 3D imaging and the degree of distortion

(Lago et al., 2021). Although the three-dimensional images captured in the wild under some special conditions can be said to be distortion-free, their three-dimensional effect is not perfect. Therefore, it is not appropriate to use non-reference evaluation methods when evaluating the visual quality of 3D images. This article studies how to use the characteristics of the human eye to evaluate users' subjective feelings from a visual perspective. Figure 1 shows the basic process of this algorithm (picture quoted from Shen et al., [2023]). The physical image here is presented in two views: left and right.

2.1. Single Shot Detection algorithm based on deep learning

The Single Shot Detection method can not only detect a single object but also realise the end-to-end detection of the object (Bakaev et al., 2023). This method has higher detection speed and efficiency than convolutional neural networks. The single-frame image detection method includes two levels: the underlying network and edge nodes. The auxiliary layer is a convolutional neural network with a decreasing scale, and the detected objects are detected simultaneously at multiple levels (Li et al., 2024). And the elements of perceptual characteristics at each level are also different. Single-frame image detection methods utilise multiple levels of feature maps. Default boxes and sensing areas were set at multiple scales. The dimensions of the default box at each level are represented in Equation (1).

$$R_t = R_{\min} + \frac{R_{\max} - R_{\min}}{n-1}(t-1) . \quad (1)$$

In Equation (1), R_t is the default box size corresponding to the t level characteristic graphic. R_{min} and R_{max} represent the minimum and maximum dimensions of the default box. The value of n depends on the single-frame detection algorithm. It represents the height of the default box in Equation (2) (Marvasti-Zadeh et al., 2021).

$$g_t^{\xi_c} = \frac{R_t}{\sqrt{\xi_c}} . \quad (2)$$

In Equation (2), ξ_c is the default value set by the single shadow detection algorithm for the aspect ratio of the preset box. $g_t^{\xi_c}$ is the height of the default box. It represents the width of the default box in Equation (3).

$$\zeta_t^{\xi_c} = R_t\sqrt{\xi_c} . \quad (3)$$

In Equation (3), $\zeta_t^{\xi_c}$ is the width of the preset box. additional default boxes (Equation [4]) are added to the default box R_t.

$$R_t' = \sqrt{R_t + R_{t+1}} . \quad (4)$$

It can be seen from Equation (4) that the characteristic graph will have six default boxes. After obtaining the characteristic map of the object, a convolution operation is performed on the object to obtain the identification and detection results of the object (Zhou et al., 2022). The detected values are divided into class confidence and bounding box localisation. And the same convolution with all classification confidences and bounding box positions is conducted. The sum of the weights of the bounding box positioning error and the confidence positioning error is called the loss function. Equation (5) represents the loss function.

$$H(x,z,h,f) = \frac{1}{M}(H_{conf}(x,z) + \xi H_{loc}(x,h,f)) . \quad (5)$$

In Equation (5), M is the forward sampling number of the prior box in the SDD algorithm. z is the numerical value of the expected classification confidence. H is the position prediction value of the bounding box corresponding to the prior box. f is the location parameter of the geographical path. According to Equation (5), the target object can be detected, and corresponding results can be generated (Castellano et al., 2021).

2.2. Modelling of the monocular sight distance measurement system

Monocular vision testing is a method that uses a camera to take photos of the object being tested and then uses the photos for actual measurement. Images captured by cameras carried by patrol robots are extracted. The camera's imaging is usually divided into two parts: lens imaging and hole imaging. Their working principle is shown in Figure 2.

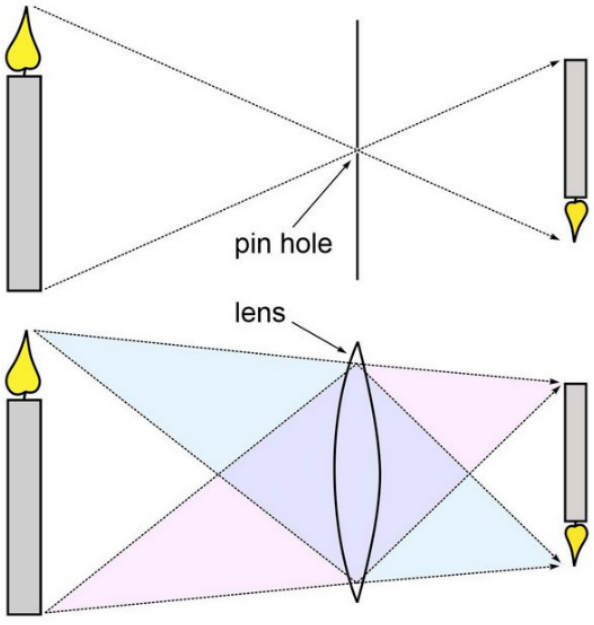

Figure 2. Lens imaging and pinhole imaging.

The projection equation of the camera in the pinhole image is derived using the principle of geometric similarity. This result is given in Equation (6).

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix}=\frac{p}{z_z}\begin{pmatrix}x_z\\ y_z\end{pmatrix}=p\begin{pmatrix}x_{zn}\\ y_{zn}\end{pmatrix}. \tag{6}$$

From Equation (6), the orientation of the inspection robot's camera and the lateral distance between obstacles can be calculated. The lateral distance is expressed in Equation (7).

$$S=(Z\times P)/Q\ . \tag{7}$$

In Equation (7), S is the lateral distance between the position of the inspection robot's camera and the obstacle. Z represents the width of the obstruction to be measured (Zhang et al., 2021). Q represents the width of the obstacle in the image. P is the focus of the position when the patrol camera is mounted. The mutual transformation of the image coordinate system, imaging plane coordinate system, camera coordinate system, and sky and earth coordinate system is realised through the transformation of the image coordinate system and the sky and earth coordinate system. Equation (8) expresses the relationship between image and imaging plane coordinates.

$$\begin{pmatrix}z\\ c\\ 1\end{pmatrix}=\begin{bmatrix}p_x & p_s & z_x\\ 0 & p_y & z_y\\ 0 & 0 & 1\end{bmatrix}\begin{pmatrix}x_{zn}\\ y_{zn}\\ 1\end{pmatrix}. \tag{8}$$

Equation (9) represents the internal parameter matrix of the camera.

$$N_{in}\begin{bmatrix}p_x & p_s & z_x\\ 0 & p_y & z_y\\ 0 & 0 & 1\end{bmatrix}. \tag{9}$$

δ is a non-zero-degree factor and it is made consistent with $\delta=1/z$; then, the transformation method from the obstacle image coordinate system to the world coordinate system is given by Equation (10).

$$\begin{pmatrix}z\\ c\\ 1\end{pmatrix}=\delta N_{in}N_{ex}\begin{pmatrix}x_\zeta\\ y_\zeta\\ z_\zeta\\ 1\end{pmatrix}. \tag{10}$$

The camera's internal parameters can be expressed by Equation (11).

$$B=\begin{bmatrix}a_x & f & \alpha_0\\ 0 & a_y & \beta_0\\ 0 & 0 & 1\end{bmatrix}. \tag{11}$$

The orthogonality between the two azimuth vectors can be obtained from the optical centre on the image axis and the two azimuth vectors. The orthogonal relationship is expressed in Equation (12).

$$\det[G]\neq 0\ . \tag{12}$$

Then G can be expressed using Equation (13).

$$G=\min\sum_i n_i-n_i^2\ . \tag{13}$$

In Equation (13), n is the coordinate of the calibration point in the image coordinate system. n' is the coordinate of the calibration point on the calibration plate (Liu et al., 2022). The limits of the camera's internal parameters can be obtained using the orthogonal relationship between the rotation matrix R and Equation (13). Finally, the internal parameters of the camera are found through restrictions. After obtaining the internal parameters, put the matrix into Equation (14) to obtain the external parameters.

$$\begin{cases}c_1=\eta Bg_1\\ c_2=\eta Bg_2\\ c_3=c_1\times c_2\\ t=\eta Bg_3\end{cases}. \tag{14}$$

B is the matrix used to find the internal parameters. η is consistent with Equation (15).

$$\eta=\frac{1}{Bg_1}=\frac{1}{Bg_2}\ . \tag{15}$$

The camera is calibrated after determining the two parameters inside and outside the camera.

3. Results

However, the current research on stereoscopic image quality assessment at home and abroad is still in its infancy. The 3D image quality assessment database built by the Multimedia Information Processing Group (MMSPG) of the Polytechnic University of Lausanne (EPFL) in Switzerland is currently the only database that can be used for 3D imaging quality assessment (Alvarado, 2022). The database was constructed to eliminate all types of distortion and be of high quality. That is to say, the change in subjective quality score caused by the degradation of image quality can be reduced to a large extent. All 3D stereoscopic

images in the EPFL library come from two high-definition cameras of the same model. The optical axes of the camera lenses are parallel to each other, ensuring that the distance between the two camera lenses is constantly adjusted. The adjustment range is 7 cm to 50 cm. Parallax can be achieved by adjusting the lateral distance between the two cameras. Due to changes in the lateral distance between the cameras, differences in the subjective ratings of stereoscopic images are caused. This method evaluates multiple stereo images in the EPFL library based on disparity. After obtaining the final target score, the quadratic index fitting formula predicts the relationship between the objective and subjective quality scores (MOS). Its expression is:

$$PredictedScore = \alpha^* \exp(b^* MOS) + c^* \exp(d^* MOS) . \quad (16)$$

This article then compares the interrelationships between subjective and objective scores. This article uses three evaluation criteria: Pearson correlation, Spearman rank correlation coefficient (SROCC) and RMSE (Jiao et al., 2021). Table 1 shows the evaluation indices of some perceptual quality evaluation algorithms on the EPFL database. It can be seen from Table 1 that the algorithm described in this article has a higher Pearson correlation and a smaller RMSE. This shows high forecast accuracy. In addition, the larger the correlation coefficient between the Spearman rank and phase, the stronger the monotonicity of the prediction. Table 1 also compares the evaluation methods without adding specific HVS features. Here, $PS_{nosaliency}$ refers to the perceptual evaluation algorithm without adding saliency detection. $PS_{\sin glescale}$ and $PS_{multiscale}$ are assessment methods on single and multiple scales. Compared with the other two methods, $PS_{multiscale}$ has a higher Pearson correlation coefficient and Spearman rank correlation coefficient and a more negligible root mean square difference. Saliency detection and multi-scale analysis are added to improve the accuracy and monotonicity of the model.

Table 1. Evaluation indicators.

Metrics	PCC	SROCC	RMSE
$PS_{nosaliency}$	0.691	0.739	0.089
$PS_{\sin glescale}$	0.748	0.751	0.076
$PS_{multiscale}$	0.785	0.755	0.071

The scatter plot shown in Figure 3 illustrates the correlation between the target score obtained by the algorithm given in this article and the subjective score obtained by the tester's score. Each point on the speckle diagram in the EPFL visual quality evaluation database represents the corresponding image. The horizontal axis in the scatter plot is the objective score obtained by the evaluation method proposed in this study, and the vertical axis is the subjective score of the Multimedia Information Processing Department of the University of Technology in Lausanne (Bayoudh et al., 2022). It can be seen from Figure 3 that the vertical axis and the horizontal axis have a high degree of correlation, which also shows that the method that the paper proposes is effective. However, it can be seen from Figure 3 that there is very little data in the chart, so in a sense, the experimental results are not convincing. This is also related to the current research status, and the database for evaluating the perceptual quality of stereoscopic images still needs to be improved.

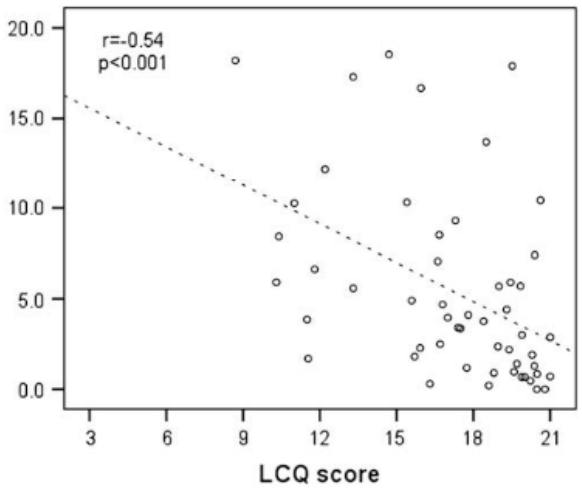

Figure 3. Comparison of objective scores and subjective scores.

4. Discussion

Judging from the current research results, the research on the visual information of images covers many fields, such as psychology, brain physiology, colour representation and emotion, line representation and emotion, etc. The quantitative factors of visualisation are challenging to quantify accurately, but fuzzy functions can characterise themes. The current international research on image visualisation mainly includes the following contents: (1) The visual information characteristics research object is the Visual Information Processing Laboratory at Florence College in Italy (Ghosh et al., 2023). The representative works have comprehensively reported the research results from 1996 to 1999. This project extracts visual information from images and videos at the semantic level. It is quantitatively analysed using the quantitative method of fuzzy sets. Then, it is applied to information representation and retrieval of text (Russell et al., 2021). At present, scholars at home and abroad have done a lot of research on the visual information characteristics of images, but they have not defined the information in them; and (2) With the continuous development of multimedia, the birth of MPEG-4 and MPEG-7 standards promoted the birth of the MPEG-7 standard. Some scholars have defined

MPEG-7 as an interface for describing multimedia information. This describes the characteristic parameters of each level, such as audio and images. The advent of MPEG-7 will significantly promote the development of the visualisation field. Existing research reports have pointed out that subjective judgement factors are introduced into video sequences to improve the video compression effect, and line structure extraction methods based on visual characteristics are introduced (Tung et al., 2022). It shows that objective evaluation and analysis of video content are essential. Eye disparity information is the most crucial external factor in human three-dimensional space. If the left and right viewpoint images are of high quality, but the binocular parallax is not correctly matched, the three-dimensional sense of the visual experience will be significantly reduced. Therefore, it is imperative to construct a three-dimensional evaluation method. This paper uses the absolute difference map method to perform three-dimensional reconstruction. The research results show that the left and right viewpoint images contain rich similarities, especially the most significant visual differences at the boundaries of the target. Therefore, in the visual system, the image's rough outline is the absolute difference between the image of a pair of viewpoints and the standard parallax of the human eye. When the absolute value difference between the distorted image pair and the absolute value difference between the original image pair is more minor, the three-dimensional effect of the image is better.

5. Conclusion

This project planned first to obtain the disparity and then solve the edge and saliency maps. Then, the weight is adjusted. Finally, Minkowski fusion technology integrates the final disparity images into a single visual score. A multi-scale analysis method calculates the visual perception scores at each level and adds the weights. This obtains the final visual quality score. An experimental study of the proposed visual effect evaluation method was conducted using EPFL. A scatter plot of subjective and objective scores was obtained, analysing each criterion. This verifies the correctness of the method. Research has found that images' structural boundaries and salient areas play a more prominent role in stereoscopic vision generation than in other areas. In addition, multi-scale analysis methods can also improve the accuracy of visual quality assessment methods.

Notes on contributors

Mr Jinsong He was born in Yue Yang, Hunan, People's Republic of China, in 1988. He received his undergraduate and master's degrees from Beijing Institute of Graphic Communication, People's Republic of China. He studied the art of photography during his undergraduate and master's degrees. Now, he is studying the combination of visual arts and traditional Chinese culture. His research analyses the ways of combining visual arts and traditional Chinese culture and their expressions from the perspective of technological iteration, and seeks a more rational and objective research method.

Dr Syamsul Bahrin Zaibon currently is an Associate Professor of Multimedia at the School of Creative Industry Management & Performing Arts, Universiti Utara Malaysia (UUM) and teaches various courses in the areas of creative industry and interactive media. Dr Syamsul holds an undergraduate degree in Information Technology from Universiti Utara Malaysia (UUM), an MSc Multimedia & Internet Computing from Loughborough University, and a PhD in Multimedia from UUM. His research is diverse and focuses on multimedia & mobile applications, web design & development, game-based learning, comics for learning, and edutainment. He has a number of research outputs and publications in these areas and presented papers at national and international conferences. He is affiliated with several professional societies (i.e., Internet Society Malaysia, IEEE) and has been the recipient of various awards in his fields of expertise, including the Geneva Exhibition & Invention, Seoul International Invention Fair, PECIPTA, Malaysian Technology Expo, and Innovation & Technology Expo. In addition, one of his journal articles was awarded as the best article for the Malaysian Journal of Learning and Instruction.

References

[1] Alvarado R (2022). Should we replace radiologists with deep learning? *Pigeons, error and trust in medical AI. Bioethics*. 36(2), pp. 121-133.

[2] Bakaev M, Heil S, Khvorostov V, and Gaedke M (2023). How Many Data Does Machine Learning in Human–Computer Interaction Need: Re-Estimating the Dataset Size for Convolutional Neural Network-Based Models of Visual Perception. *IT Professional*. 25(2), pp. 23-29.

[3] Bayoudh K, Knani R, Hamdaoui F, and Mtibaa A (2022). A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*. 38(8), pp. 2939-2970.

[4] Bonnen T, Yamins DL, and Wagner AD (2021). When the ventral visual stream is not enough: A deep learning account of medial temporal lobe involvement in perception. *Neuron*. 109(17), pp. 2755-2766.

[5] Castellano G, and Vessio G (2021). Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. *Neural Computing and Applications*. 33(19), pp. 12263-12282.

[6] Funke CM, Borowski J, Stosio K, Brendel W, Wallis TS, and Bethge M (2021). Five points to check when comparing visual perception in humans and machines. *Journal of Vision*. 21(3), pp. 16-16.

[7] Ghosh S, Dhall A, Hayat M, Knibbe J, and Ji Q (2023). Automatic gaze analysis: A survey of deep learning based approaches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 46(1), pp. 61-84.

[8] Jiao L, Zhang R, Liu F, Yang S, Hou B, Li, L, and Tang X (2021). New generation deep learning for video object detection: A survey. *IEEE Transactions on Neural Networks and Learning Systems*. 33(8), pp. 3195-3215.

[9] Johnson E, and Nica E (2021). Connected vehicle technologies, autonomous driving perception algorithms, and smart sustainable urban mobility behaviors in networked transport systems. *Contemporary Readings in Law and Social Justice*. 13(2), pp. 37-50.

[10] Lago F, Pasquini C, Böhme R, Dumont H, Goffaux V, and Boato G (2021). More real than real: A study on human visual perception of synthetic faces [applications corner]. *IEEE Signal Processing Magazine*. 39(1), pp. 109-116.

[11] Li R, Cao Y, Tang H, and Kaiser G (2024). Teachers' scaffolding behavior and visual perception during cooperative learning. *International Journal of Science and Mathematics Education*. 22(2), pp. 333-352.

[12] Li X, Wu D, Li L, and Wang X (2022). Research on visual perception evaluation of urban riverside greenway landscape based on deep learning. *Journal of Beijing Forestry University*. 43(12), pp. 93-104.

[13] Liu X, Chen S, Song L, Woźniak M, and Liu S (2022). Self-attention negative feedback network for real-time image super-resolution. *Journal of King Saud University-Computer and Information Sciences*. 34(8), pp. 6179-6186.

[14] Marvasti-Zadeh SM, Cheng L, Ghanei-Yakhdan H, and Kasaei S (2021). Deep learning for visual tracking: A comprehensive survey. *IEEE Transactions on Intelligent Transportation Systems*. 23(5), pp. 3943-3968.

[15] Russell RL, and Reale C (2021). Multivariate uncertainty in deep learning. *IEEE Transactions on Neural Networks and Learning Systems*. 33(12), pp. 7937-7943.

[16] Shen L, Yao Y, Geng X, Fang R and Wu D (2023). A novel no-reference quality assessment metric for stereoscopic images with consideration of comprehensive 3D quality information. *Sensors*. 23(13), pp. 6230

[17] Tsuneki M (2022). Deep learning models in medical image analysis. *Journal of Oral Biosciences*. 64(3), pp. 312-320.

[18] Tung TY, and Gündüz D (2022). DeepWiVe: Deep-learning-aided wireless video transmission. *IEEE Journal on Selected Areas in Communications*. 40(9), pp. 2570-2583.

[19] Zhang X (2021). Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 44(9), pp. 4819-4838.

[20] Zhang Y, Liu H, Yang Y, Fan X, Kwong S, and Kuo CJ (2021). Deep learning based just noticeable difference and perceptual quality prediction models for compressed video. *IEEE Transactions on Circuits and Systems for Video Technology*. 32(3), pp. 1197-1212.

[21] Zhou Y, Xiao J, Zhou Y, and Loianno G (2022). Multi-robot collaborative perception with graph neural networks. *IEEE Robotics and Automation Letters*. 7(2), pp. 2289-2296.