# R2-B2: A Metric of Synthesized Image's Photorealism by Regression Analysis based on Recognized Objects' Bounding Box

1st Shun Hattori
*Faculty of Advanced Engineering*
*The University of Shiga Prefecture*
Hikone-shi, Japan
hattori.s@e.usp.ac.jp

2nd Kizuku Aiba
*Graduate School of Science and Engineering*
*Muroran Institute of Technology*
Muroran-shi, Japan
21043001@mmm.muroran-it.ac.jp

3rd Madoka Takahara
*Faculty of Advanced Science and Technology*
*Ryukoku University*
Otsu-shi, Japan
takahara@rins.ryukoku.ac.jp

***Abstract*—In recent years, a lot of researches on AI (Artificial Intelligence) for Image Synthesis and Image Generation have been being conducted actively, and state of the art GANs (Generative Adversarial Networks) for text-to-image have been able to generate precise images with high photorealism for a text-based user query (but also no-good images). However, it is pointed out that the precision of all images generated for a query has been not always enough high. Therefore, for practical usages, they are required to be re-ranked and/or filtered based on some sort of metric(s). This paper proposes a novel metric, R2-B2 (RR-BB), on photorealism, especially "size balance" (i.e., balance between in-image objects' size), of a manually or automatically synthesized image by Regression analysis based on multiple Recognized objects' Bounding Box, i.e., the position $(x, y)$ and size (width, height, or area) of objects recognized in the image.**

***Index Terms*—image evaluation, no-reference image quality assessment, image quality metrics, object recognition, object detection, regression analysis, correlation analysis**

## I. Introduction

In recent years, a lot of researches on AI (Artificial Intelligence) for Image Synthesis and Image Generation have been being conducted actively, and state of the art GANs (Generative Adversarial Networks) [1], [2] for text-to-image have been able to generate precise images with high photorealism for a text-based user query (but also no-good images). However, it is pointed out that the precision of all images generated for a query has been not always enough high.Therefore, for practical usages, they have been yet required to be re-ranked and/or filtered based on some sort of metric(s) [3]–[6]. This paper proposes a novel R2-B2 (RR-BB) metric on photorealism, especially "size balance" (i.e., balance between in-image objects' size) or "size sense" [7], of a manually or automatically synthesized image by Regression analysis based on multiple Recognized objects' Bounding Box, i.e., the position $(x, y)$ and size (width, height, or area) of objects recognized in the image.

Figure 1 shows no-good images without photorealism, especially size balance, generated and selected by the Parti (Pathways Autoregressive Text-to-Image) model [2], that is reported to be the SOTA (State-Of-The-Art) text-to-image with MS-COCO-FID-30K = 7.23 on July 22nd, 2022 [8]. And also Figure 1 shows their CLIP [4] score for their text-based query, that is a similarity between an image and a text-based query, and the proposed R2-B2 score for no extra input(s) such as a text-based query and a reference image [11], [12]. They have high similarity (i.e., CLIP score) for their text-based query, but seem to be low balance between in-image objects' size in the real world. Therefore, their R2-B2 score is required to be calculated to be low by the proposed method, while the R2-B2 score for generated images (a,b,c) with some failures for their query but with high size balance and a real photo image (d) in Figure 2 is required to be calculated to be high.

## II. Related Work

The proposed method to calculate the novel R2-B2 metric on "size balance" (i.e., balance between in-image object's size) of an input image is mainly related to the research fields of IQA (Image Quality Assessment) [9]–[18], IQMs (Image Quality Metrics) [20]–[22], Image Evaluation [3]–[6] for a text-based query, and Image Retrieval [23], [24]. These are very similar to each other, because IQA, IE, and IR other than IQMs also use some sort of metric(s) for an input image.

The research field of IQA, especially not subjective but objective IQA, has a lot of existing researches, and is divided into three kinds of IQA: FRIQA (Full-Reference IQA) [11], [12], RRIQA (Reduced-Reference IQA) [13]–[15], and NRIQA (Non-Reference IQA) [16], [17]. NRIQA is to assess

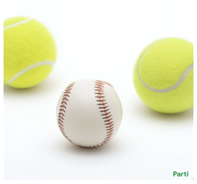
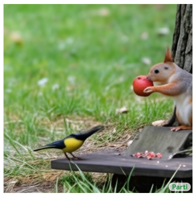
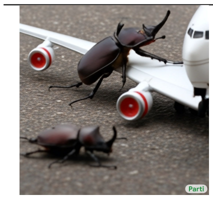
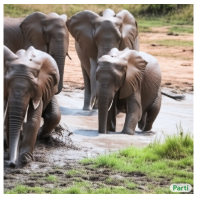
| | |
|---|---|
|  | query = “Two baseballs to the left of three tennis balls.”<br>CLIP score = 32.0625 (high for the query)<br>R2-B2 score = ? (to be low because a baseball is nearer than two tennis balls but a baseball is smaller than two tennis balls.)<br>in-image objects’ size in real world:<br>baseball = 73 to 76 [mm] in diameter<br>tennis ball = 65.41 to 68.58 [mm] in diameter |
|  | query = “A squirrel gives an apple to a bird.”<br>CLIP score = 34.4375 (high for the query, while low for the other 3 queries.)<br>R2-B2 score = ? (to be low because a squirrel and a bird seem to be big for an apple.)<br>in-image objects’ size in real world:<br>squirrel = 10 to 127 [cm] in total length<br>apple = 5.5 to 8.5 [cm] in diameter<br>bird = 5.5 to 280 [cm] in total length |
|  | query = “A rhino beetle this size of a tank grapples a real life passenger airplane on the tarmac.”<br>CLIP score = 34.8125 (high for the query, while low for the other 3 queries.)<br>R2-B2 score = ? (to be low because two rhino beetles are too big for an airplane.)<br>in-image objects’ size in real world:<br>rhino(ceros) beetle = 40 to 80 [mm] in total length<br>airplane (jet) = 33.6 to 73.9 [m] in total length |
|  | query = “A group of elephants walking in muddy water.”<br>CLIP score = 34.0625 (high for the query, while low for the other 3 queries.)<br>R2-B2 score = ? (to be low? because an elephant (not child) centered in the image is nearer than two elephants following it, but it seems to be smaller than them.)<br>in-image objects’ size in real world:<br>elephant (adult) = 247 to 336 [cm] in height<br>elephant (baby) = about 100 [cm] in height |

Fig. 1. Examples of no-good images generated by the SOTA text-to-image, Parti [2], and their CLIP [4] score and the proposed R2-B2 score, where information about in-image objects’ size in the real world is used Wikipedia as a reference.

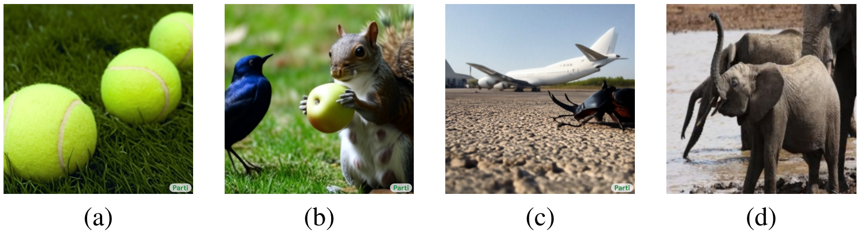


Fig. 2. Examples of generated images (a,b,c) with higher balance between in-image objects’ size than the images in Figure 1 and real photo images (d).

the quality [1] of an input image for no extra input(s) such as a reference image and a text-based query, like the proposed R2-B2. But IQA including NRIQA is to assess not “sense balance” but 25 types of distortions such as blur, color diffusion and saturation, JPEG compression, and various noises [10], unlike the proposed R2-B2.

Image Evaluation [3]–[6] and Image Retrieval [23], [24] are to calculate some sort of metric(s) such as the fitness and a similarity of an input image for a text-based query and/or a reference image, and to offer a user images ranked or filtered based on the metric(s). Fundamentally, they require extra input(s), unlike the proposed R2-B2 and NRIQA.

Figure 3 gives an overview of the proposed R2-B2 metric and its related research fields with focusing on Input/Output. Aiba, et al. [26] proposes a support system for Image Synthesis by estimating the spatial structure of a background image using regression analysis based on recognized objects’ position and size in a background image and automatically adjusting the size of a pasted object image to synthesize images with “size balance,” while Nishihara, et al. [7] proposes a 3DCG CAPTCHA system using arbitrarily-distorted images without “size sense” between in-image objects. Layout-to-Image GANs [27]–[29] generate images adapted to not only

[1] Image Quality is defined as “the level of accuracy with which different imaging systems capture, process, store, compress, transmit and display the signals that form an image” in Wikipedia [19] from a perspective of signal processing systems or “the weighted combination of all of the visually significant attributes of an image” [21] from a perspective of human viewers.

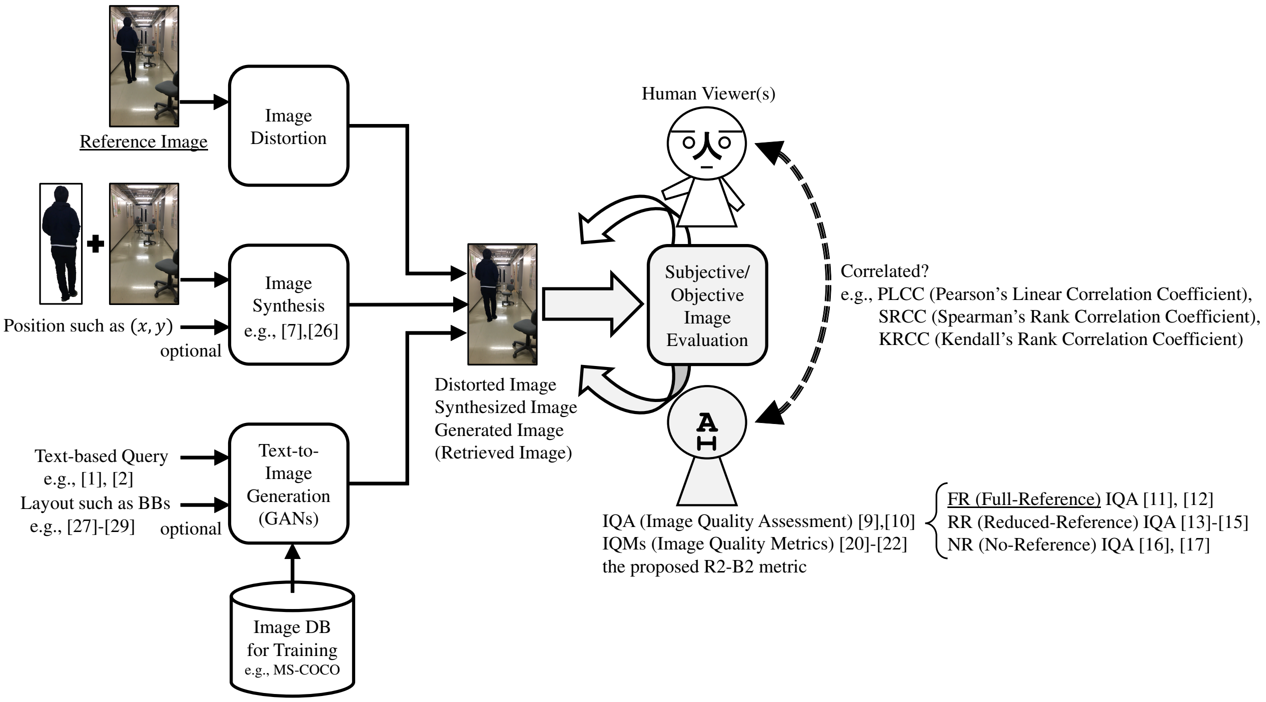

Reference Image
Image Distortion
Image Synthesis e.g., [7],[26]
Position such as $(x, y)$
optional
Text-based Query e.g., [1], [2]
Layout such as BBs e.g., [27]-[29]
optional
Text-to-Image Generation (GANs)
Image DB for Training e.g., MS-COCO
Distorted Image
Synthesized Image
Generated Image
(Retrieved Image)
Human Viewer(s)
Subjective/ Objective Image Evaluation
Correlated?
e.g., PLCC (Pearson's Linear Correlation Coefficient),
SRCC (Spearman's Rank Correlation Coefficient),
KRCC (Kendall's Rank Correlation Coefficient)
IQA (Image Quality Assessment) [9],[10]
IQMs (Image Quality Metrics) [20]-[22]
the proposed R2-B2 metric
FR (Full-Reference) IQA [11], [12]
RR (Reduced-Reference) IQA [13]-[15]
NR (No-Reference) IQA [16], [17]

Fig. 3. An overview of the proposed R2-B2 metric and its related research fields with focusing on Input/Output.

a text-based query but also the layout of in-image objects such as their BBs (Bounding Boxes), but do not evaluate "size balance" (i.e., balance between in-image objects' size), unlike the proposed R2-B2 metric.

## III. Proposed Method

This paper proposes a novel R2-B2 metric on photorealism, especially "size balance" (i.e., balance between in-image objects' size), of a manually or automatically synthesized image by Regression analysis based on multiple Recognized objects' Bounding Box, i.e., the position $(x, y)$ and size (width, height, and/or area) of objects recognized in the image.

This section first gives an overview of the idealized system and the realized system for the proposed R2-B2 metric, and also described the limitations of the current R2-B2 metric (e.g., the requirements of input images). Second, the algorithm to calculate the proposed R2-B2 metric for an input image with fulfilling the requirements is defined in detail. Finally, several existing metrics to validate the proposed R2-B2 metric by comparing with them in the experiments are introduced.

### A. Overview and Limitations

Figure 4 gives an overview of the idealized system(s) and the realized system for the proposed R2-B2 metric.

The current R2-B2 metric has several limitations and input images to be evaluated might have some requirements:

- The proposed algorithm can reflect not any instance/subclass and its doing/transforming by a specific/generic object recognition with high precision and recall, but only the classes that can be recognized by a generic object recognition API used in the proposed algorithm's Step 1. This paper uses TensorFlow 2 Object Detection API [30] with the pre-trained model, CenterNet HourGlass104 1024x1024, that can recognize 80 classes with MS-COCO mean Average Precision = 44.5 [31].
- The proposed algorithm can utilize the real-world size of a recognized object in an input image to convert the object to the Unit Object in the proposed algorithm's Step 2, not enough fine-grained, i.e., instance-level and specialized for a user (human viewer) (e.g., S. Hattori), a space (e.g., in Japan), and a time (e.g., in 2022) by using the spatiotemporally-localized Object DBs, but only coarse-grained, i.e., class-level and ad-hoc generalized by using the generalized Object DB shown in Table I.
- Input images to be evaluated might have some requirements: with vanishing point(s) (while with/out vanishing lines), and without rotated, especially by a rotation angle other than 90°, 180°, and 270° (while with/out reversed).

### B. Definition of R2-B2 Metric

The proposed R2-B2 algorithm to evaluate the "size balance" (i.e., balance between in-image objects' size) of an input image without extra input(s) such as a text-based query and a reference image is defined as follows.

input: an image $img$ (that meets requirements)
and ideally, a user (human viewer) $u$ (that is optional)
output: the R2-B2 score, $img$.R2-B2$(u) \in [-1, 1]$, of the input image $img$ (ideally, personalized or not for the input user $u$, but in this paper, only not personalized)

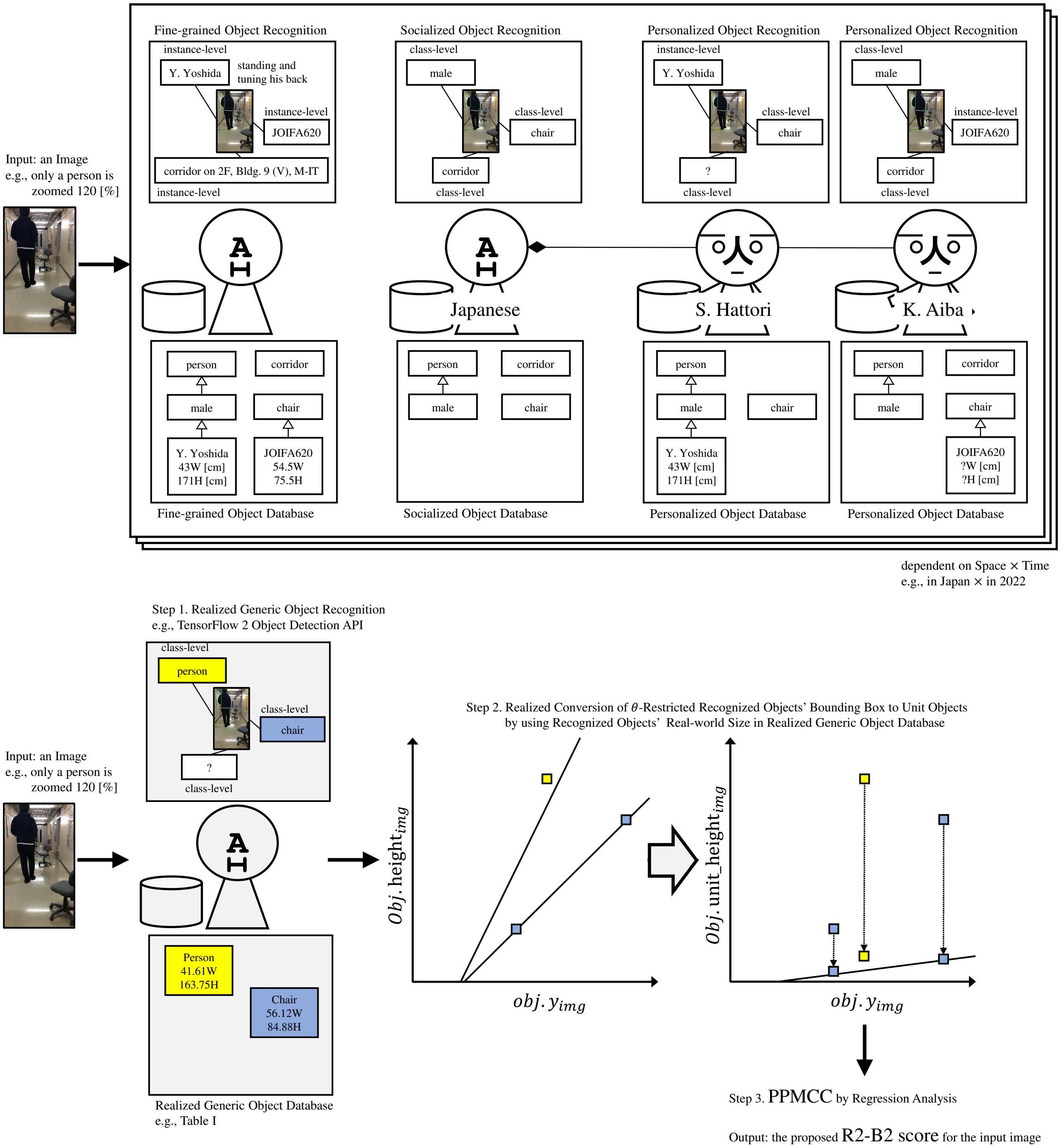

Input: an Image
e.g., only a person is
zoomed 120 [%]
Fine-grained Object Recognition
instance-level
Y. Yoshida
standing and
tuning his back
instance-level
JOIFA620
corridor on 2F, Bldg. 9 (V), M-IT
instance-level
Socialized Object Recognition
class-level
male
class-level
chair
corridor
class-level
Personalized Object Recognition
instance-level
Y. Yoshida
class-level
chair
?
class-level
Personalized Object Recognition
class-level
male
instance-level
JOIFA620
corridor
class-level
Japanese
S. Hattori
K. Aiba
person
corridor
male
chair
Y. Yoshida
43W [cm]
171H [cm]
JOIFA620
54.5W
75.5H
Fine-grained Object Database
person
corridor
male
chair
Socialized Object Database
person
male
chair
Y. Yoshida
43W [cm]
171H [cm]
Personalized Object Database
person
corridor
male
chair
JOIFA620
?W [cm]
?H [cm]
Personalized Object Database
dependent on Space × Time
e.g., in Japan × in 2022
Step 1. Realized Generic Object Recognition
e.g., TensorFlow 2 Object Detection API
class-level
person
class-level
chair
?
class-level
Input: an Image
e.g., only a person is
zoomed 120 [%]
Person
41.61W
163.75H
Chair
56.12W
84.88H
Realized Generic Object Database
e.g., Table I
Step 2. Realized Conversion of $\theta$-Restricted Recognized Objects' Bounding Box to Unit Objects
by using Recognized Objects' Real-world Size in Realized Generic Object Database
$Obj.\mathrm{height}_{img}$
$obj.y_{img}$
$Obj.\mathrm{unit\_height}_{img}$
$obj.y_{img}$
Step 3. PPMCC by Regression Analysis
Output: the proposed R2-B2 score for the input image

Fig. 4. Ideal object recognition that is instance-level, spatio-temporal, and personalize, and the realized R2-B2 algorithm using 80-class-level object recognition.

Step 1. (Object-oriendted) Object Recognition

To calculate the R2-B2 score by Regression analysis based on multiple unit objects' Bounding Box converted from Recognized objects' Bounding Box) in the Step 3, this Step 1 has to recognize in-image objects' name of class/instance and bounding box in the input image $img$, as precisely and exhaustively as possible, by not specific but genetic object recognition. This paper uses TensorFlow 2 Object Detection API [30] with the pre-trained model, CenterNet HourGlass104 1024x1024, that can recognize 80 classes (labels) with MS-COCO mAP (mean Average Precision) $= 44.5$ [31].

Step 1.5. In-image Space and Time Estimation

Ideally, the Step 1.5 estimates the space $s$ (e.g., in Japan) and the time $t$ (e.g., in 2022) in the input image, to specialize and more accurately estimate recognized objects' real-world size by using not the generalized but the spatiotemporally-localized Object DB for the space $s$ and the time $t$ in the next Step 2. However, this Step 1.5 is skipped in this paper, and would be tackled in the future work.

Step 2. Conversion of Recognized Objects to Unit Objects

To apply regression analysis unifiedly to heterogeneous classes (ultimately, instances) of recognized objects' bounding box, i.e., the position and size in the input image in the next Step 3, this Step 2 converts the in-image size, $obj.\text{size}_{img}$, of a recognized object $obj$ in the input image $img$ to the in-image size, $obj.\text{unit\_size}_{img}$, of the Unit Object (in this paper, cubes 1 [cm] on a side) by using the average size(s) $\mu$ of its class-name in the Object DB (as shown in Table I):

$$\forall obj,\ \ obj.\text{unit\_size}_{img} := \frac{obj.\text{size}_{img}}{obj.\text{size}_{rw}(u,s,t)} \tag{1}$$

where $obj.\text{size}_{img}$ stands for the field (constant value) of an object $obj$ to return the in-image size (at the position $(x, y)$) of a recognized object $obj$ in an input image $img$, and $obj.\text{size}_{rw}(u, s, t)$ stands for the method (function) of an object $obj$ to return the real-world size of a recognized object $obj$ in an input image, ideally, specialized for a user (human viewer) $u$ (e.g., S. Hattori) in a space $s$ (e.g., in Japan) and time $t$ (e.g., in 2022) by using the spatiotemporally-localized Object DB, but in this paper, generalized (not specialized for any user in any space and any time) by using the generalized Object DB. In this paper, the Object DB is generalized by averaging the average size(s) $\mu$ in the spatiotemporally-localized Object DB(s). For instance, the generalized (not specialized) average width, $obj.\text{size}_{rw}() = 41.605$ [cm], for any recognized object $obj$ of the class "person" is averaged by 43.95 [cm] for Young/Male Japanese and 39.26 [cm] for Young/Female Japanese shown in Table I.

Step 3. Regression (Correlation) Analysis

To evaluate the "size balance" of an input image without extra input(s) such as a text-based query and a reference image, this Step 3 conducts regression (correlation) analysis between the position and the in-image size of the Unit Objects, converted from all recognized objects $\forall obj$ in an input image. Here, "all" recognized objects are restricted with their recognition probability (e.g., "detection_scores" by TensorFlow 2 Object Detection API [30]) $>$ a threshold $\theta$.

In this paper, the PPMCC (Pearson's Product-Moment Correlation Coefficient) between the position and the in-image size of the Unit Objects, converted from all recognized objects $\forall obj$ in an input image is output without any change, as the proposed R2-B2 score $img.\text{R2-B2}(u) \in [-1, 1]$ to evaluate the "size balance" of an input image $img$, ideally, specialized for a user (human viewer) $u$ (e.g., S. Hattori), but in this paper, generalized (not personalized for any user). Note that if the number of the objects recognized in an input image and filtered with a threshold $\theta$ is less than 2, the proposed R2-B2 metric returns "undefined" by the PPMCC. Therefore, a threshold $\theta$ cannot be set too high.

In the future work, the proposed R2-B2 metric might be adapted with some changes for a practical usage, e.g., range conversion to $[0, 1]$ where 0 means "perfectly size-unbalanced" while 1 means "perfectly size-balanced."

### C. Definition of Baseline Metrics

As a baseline for the experiments to validate the proposed R2-B2 metric, this subsection first introduces conventional FRIQA metrics for image similarity between an input image and its most photoreal image [2]: MSE (Mean Squared Error), RMSE (Root MSE), PSNR (Peak Signal-to-Noise Ratio) [dB], and MSSIM (Mean Structural SIMilarity Index) [32].

First, the metric $\text{MSE}(I, K) \in [0, \text{MAX}_I^2]$ [3] is defined as follows, that takes two arguments: an input image $I$ and its most photoreal image $K$ with their common width $w$ and height $h$. Here, $\text{MAX}_I$ is the maximum possible pixel intensity of the input image $I$.

$$\text{MSE}(I, K) := \frac{1}{w \cdot h} \sum_{i=0}^{w-1} \sum_{i=0}^{h-1} [I(i,j) - K(i,j)]^2 \tag{2}$$

where $I(i, j)$ and $K(i, j)$ denote their pixel intensities at the in-image position $(i, j)$.

Second, the metric $\text{PSNR}(I, K) \in [0, \infty]$ [4] is defined as follows, that takes 2 arguments: an input image $I$ and its most photoreal image $K$ with their common width $w$ and height $h$.

$$\text{PSNR}(I, K) := 10 \cdot \log_{10} \left( \frac{\text{MAX}_I^2}{\text{MSE}(I, K)} \right) \tag{3}$$

The metric $\text{MSSIM}(I, K) \in [0, 1]$ [5] computes the Mean SSIM (Structural SIMilarity) index between two images [33]: an input image $I$ and its most photoreal image $K$. $\text{SSIM}(x, y)$ uses raw pixel intensities locally with two sliding windows $x$ and $y$ of common size $N \times N$, while MSE and PSNR uses raw pixel intensities globally and have limitation(s).

[2] In practice, the most photoreal (ground truth) image for an input image is not always given. In this paper, it is manually given only for the experiments.

[3] 0 indicates the perfect similarity. The greater MSE, the lesser similarity.

[4] $\infty$ (or undefined) indicates the perfect similarity. The lesser PSNR, the lesser similarity.

[5] 1 indicates the perfect similarity, while 0 indicates no similarity.

TABLE I

6 (WITH THRESHOLD $\theta = 0.15$ IN THE IMAGES FOR EXPERIMENTS) OF 80 RECOGNIZABLE CLASSES BY TENSORFLOW 2 OBJECT DETECTION API [30] AND THE STATISTICS ON THEIR SIZE IN THE OBJECT DB.

| id | class-name | user | space | time | width $\mu \pm \sigma$ [cm]<br>height $\mu \pm \sigma$ [cm] | #samples | samples from |
|---|---|---|---|---|---|---|---|
| 1 | person | — | in Japan | in 2003 | width: $43.95 \pm 2.66$<br>height: $169.67 \pm 6.54$ | 49 | Young/Male in AIST/HQL 3D Anthropometric DB '03 [37] |
| | | | | | width: $39.26 \pm 1.60$<br>height: $157.83 \pm 4.35$ | 47<br>48 | Young/Female in AIST/HQL 3D Anthropometric DB '03 [37] |
| 44 | bottle | — | in Japan | in 2022 | width: $7.14 \pm 2.04$<br>height: $21.89 \pm 6.20$ | 76 | noise-filtered from the top 100 products by Amazon.co.jp [38] |
| | | | in World | | width: $7.92 \pm 2.62$<br>height: $24.44 \pm 6.24$ | 19 | 19 subclasses of bottle searched by Dimensions.com [39]<br>e.g., Soda Bottle - 2 Liter (11.0Diameter × 31.5H [cm]) |
| 62 | chair | — | in Japan | in 2022 | width: $52.58 \pm 12.63$<br>height: $84.91 \pm 28.74$ | 80 | noise-filtered from the top 100 products by Amazon.co.jp [38] |
| | | | in World | | width: $59.66 \pm 17.61$<br>height: $84.84 \pm 10.12$ | 29 | 29 subclasses of chair searched by Dimensions.com [39]<br>e.g., CH36 Chair (48.0D × 52.1W × 81.0H [cm]) |
| 67 | dining table | — | in Japan | in 2022 | width: $113.70 \pm 33.14$<br>height: $72.79 \pm 4.20$ | 93 | noise-filtered from the top 100 products by Amazon.co.jp [38] |
| | | | in World | | width: $166.73 \pm 58.44$<br>height: $73.20 \pm 2.32$ | 27 | 27 subclasses of dining table searched by Dimensions.com [39]<br>e.g., Marais Dining Table (129.5Length × 69.9W × 73.7H [cm]) |
| 72 | tv | — | in Japan | in 2022 | width: $99.78 \pm 32.38$<br>height: $63.34 \pm 19.15$ | 80 | noise-filtered from the top 100 products by Amazon.co.jp [38] |
| | | | in World | | width: $143.97 \pm 25.67$<br>height: $89.52 \pm 15.55$ | 6 | 6 subclasses of dining table searched by Dimensions.com [39]<br>e.g., Samsung 82" Q70 TV (38.6D × 183.4W × 114.6H [cm]) |
| 84 | book | — | in Japan | in 2022 | width: $18.45 \pm 4.04$<br>height: $24.25 \pm 5.25$ | 50 | noise-filtered from the top 100 products by Amazon.co.jp [38] |
| | | | in World | | — | 0 | no subclass of book searched by Dimensions.com [39] |

All the above-mentioned baseline metrics are FRIQAs, thus are not for the "size balance" (i.e., balance between in-image objects' size) but only for the "quality" of an input image, and require not only the input image but also its reference image, i.e., the most photoreal (ground truth) image. Therefore, to validate the proposed R2-B2 metric without any reference image, BRISQUE (Blind/Referenceless Image Spatial QUality Evaluator) [34] is also introduced as a NRIQA, that operates in the spatial domain based on a NSS (Natural Scene Statistic) to quantify possible losses of "naturalness" in an input image. This paper uses all the above-mentioned baseline metrics implemented by OpenCV [35] for the experiments.

The alternative image similarity methods based on keypoint detectors and local invariant descriptors are SIFT, SURF, KAZE, AKAZE, ORB, BRISK, and so forth [36]. Furthermore, there are image similarity methods based on DL (Deep Learning), especially Siamese Networks [18].

## IV. EXPERIMENTAL RESULTS

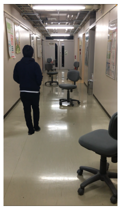
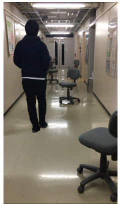
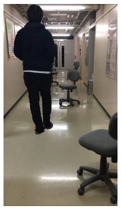
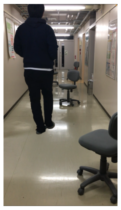
This section shows some experimental results to validate the proposed R2-B2 metric by comparing with three FRIQA metrics, MSE, PSNR, and MSSIM, and one NRIQA metric, BRISQUE, implemented by OpenCV [35].

First, Table II compares their scores for input images: a reference image which is "1 person (who is Y. Yoshida [40] with about 43W × 171H [cm]) and 5 chairs (which is the same type, JOIFA620-287F, with about 54.5W × 75.5H [cm]) in a corridor (which is on the 2nd floor of Education and Research Building 9 (V), Muroran Institute of Technology)" as the most photoreal (ground truth) image for these input images, and its distorted images by zooming only the person with 80 to 120 percent but not switching its pixel aspect ratio.

Second, Figure 5 compares the proposed R2-B2 scores (by regression-analyzing $obj.y_{img} \rightarrow obj.\text{unit\_height}_{img}$, optimizing its threshold $\theta$ and/or assuming not 80-class-level but instance-level object recognition) with the BRISQUE score as a NR/Blind IQA per zoom [%], and shows that the proposed R2-B2 metric (if by instance-level object recognition) has the peak, 0.999, at zoom 100 [%] that is not distorted, which is almost ideal, while the proposed R2-B2 metric (only by optimizing its threshold $\theta$) unfortunately has the peak, 0.993, at zoom 88 [%] that is slightly distorted, and the existing BRISQUE metric does not peak out, and also that both the proposed R2-B2 metrics fall down steeply for zoom 116 to 120 [%] that is remarkably distorted enough to make the person's head peek out in the image.

Finally, Figure 6 shows the dependency of SRCC (Spearman's Rank Correlation Coefficient) between the proposed R2-B2 score (by 80-class-level TensorFlow 2 Object Detection API [30]) and zoom [%] on its threshold $\theta$, while Figure 7 shows the dependency of SRCC (Spearman's Rank Correlation Coefficient) between the proposed R2-B2 score (if by instance-level object recognition) and zoom [%] on its threshold $\theta$. Figure 6 shows that the proposed R2-B2 metric (implemented by 80-class-level object recognition) fortunately achieves almost $-1.0$ on SRCC for zoom 100 to 120 [%], while unfortunately not near 1.0 but rather negative on SRCC for zoom 80 to 100 [%], while Figure 7 shows that the proposed R2-B2 metric (if by instance-level object recognition) can achieve almost 1.0 on SRCC for zoom 80 to 100 [%] and almost $-1.0$ on SRCC for zoom 100 to 120 [%]. Note that the existing BRISQUE metric achieves 0.996 (not negative) on SRCC for zoom 100 to 120 [%] and 1.0292 on MAE (Mean Absolute Error).

TABLE II
COMPARISON BETWEEN THE PROPOSED R2-B2 SCORE FOR "SIZE BALANCE" AND BASELINE METRICS FOR NR- OR FR-IQA.

| | | | | | | Reference Image | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | |  | | | |  | | |  |  |
| | Zoom [%] | 80 | 85 | 90 | 95 | 100 | 105 | 110 | 115 | 120 |
| FR- | MSE | 893.2714 | 739.595 | 526.237 | 298.255 | 0.000 | 311.835 | 565.195 | 895.776 | 1244.088 |
| | PSNR [dB] | 18.621 | 19.441 | 20.919 | 23.385 | $\infty$ | 23.192 | 20.609 | 18.609 | 17.182 |
| | MSSIM | 0.924 | 0.934 | 0.946 | 0.961 | 1.000 | 0.959 | 0.940 | 0.922 | 0.905 |
| NR- | BRISQUE | 24.104 | 24.206 | 24.202 | 25.125 | 25.615 | 26.524 | 27.828 | 28.812 | 29.621 |
| | R2-B2 (optimized) | 0.982 | 0.983 | 0.989 | 0.992 | 0.988 | 0.982 | 0.974 | 0.965 | 0.866 |

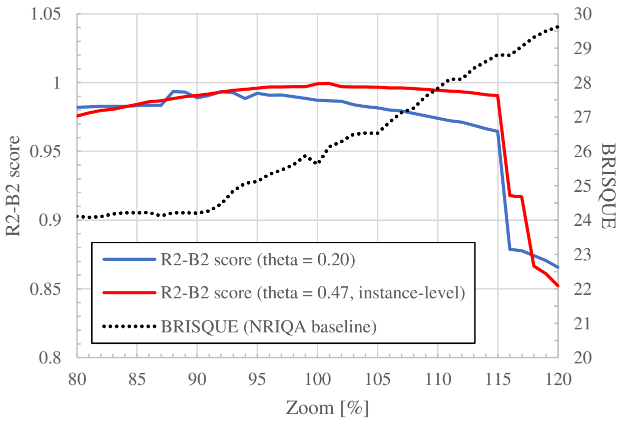


Fig. 5. Comparison between the proposed R2-B2 scores (by optimizing $\theta$ and/or assuming instance-level object recognition) and BRISQUE per zoom.

## V. CONCLUSIONS AND FUTURE WORK

To evaluate and re-rank and/or filter synthesized images based on not only a similarity between an image and its text-based query such as CLIP [3] but also "size balance" (i.e., balance between in-image objects' size), this paper has proposed a novel R2-B2 (RR-BB) metric on photorealism, especially "size balance," of a manually or automatically synthesized image by Regression analysis based on multiple Recognized objects' Bounding Box, i.e., the position $(x, y)$ and size (width, height, and/or area) of objects recognized in the image. The experimental results have shown the potentials of the proposed R2-B2 metric for "size balance" of input images, and also its limitations, especially, by not instance-level but only 80-class-level object recognition.

In the future work, the proposed R2-B2 (Regression analysis based on multiple Recognized objects' Bounding Box in an input image) metric has to be validated for as many input images as possible by various correlation coefficients with human viewers' evaluation such as MOS/DMOS (Difference Mean Opinion Score), would be personalized for a user (human viewer) and/or spatiotemporally-localized for a space and a time, and could be applied to

- object's size error detection and correction, e.g., for incorrect perspective in a frame (i.e., "koma") of a "manga,"
- vanishing point detection not based on vanishing lines,
- horizon detection and/or rotation angle estimation, only for real photo images, and so forth.

## REFERENCES

[1] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," arXiv:2205.11487, May 2022.

[2] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z.R. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu, "Scaling Autoregressive Models for Content-Rich Text-to-Image Generation," arXiv:2206.10789, June 2022.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," Proceedings of the 38th International Conference on Machine Learning (PMLR'21), vol. 139, pp. 8748–8763, July 2021.

[4] openAI, "CLIP (Contrastive Language-Image Pre-Training)," https://github.com/openai/CLIP, July 2022.

[5] T. M. Dinh, R. Nguyen, and B.-S. Hua, "TISE: A Toolbox for Text-to-Image Synthesis Evaluation," arXiv:2112.01398v1, December 2021.

[6] S. Hattori, and K. Aiba, "GANs-output Image Filtering Based on Image Evaluation Using Multiple CNNs," Image Laboratory, vol. 33, no.5, pp. 47–55, Japan Industrial Publishing, April 2022.

[7] D. Nishihara, and Ismail Arai, "Evaluation of 3D CG CAPTCHA Using Object Size Sense," IPSJ SIG Technical Report, vol. 2017-DPS-170/2017-CSEC-76, no. 4, pp. 1–8, March 2017.

[8] Papers with Code, "Text-to-Image Generation on COCO," https://paperswithcode.com/sota/text-to-image-generation-on-coco, July 2022.

[9] Y. Niu, Y. Zhong, W. Guo, Y. Shi, and P. Chen, "2D and 3D Image Quality Assessment: A Survey of Metrics and Challenges," *IEEE Access*, vol. 7, pp. 782–801, December 2018.

[10] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A Large-scale Artificially Distorted IQA Database," Proceedings of the Tenth International Conference on Quality of Multimedia Experience (QoMEX'19) pp. 1–3, June 2019.

[11] M. Pedersen, and J. Y. Hardeberg, "Full-Reference Image Quality Metrics: Classification and Evaluation," *Foundations and Trends in Computer Graphics and Vision*, vol. 7, no. 1, pp. 1–80, March 2012.

[12] K. Ding, K Ma, S. Wang, and E. P. Simoncelli, "Comparison of Full-Reference Image Quality Models for Optimization of Image Processing Systems," International Journal of Computer Vision, vol. 129, pp. 1258–1281, January 2021.

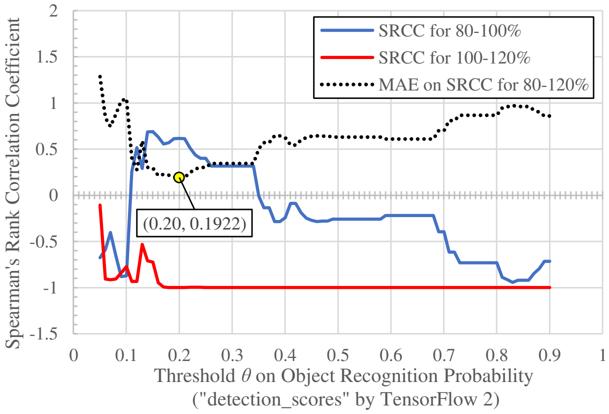


Fig. 6. The dependency of SRCC (Spearman's Rank Correlation Coefficient) between the proposed R2-B2 score (by TensorFlow 2 Object Detection API [30]) and zoom [%] on its threshold $\theta$.

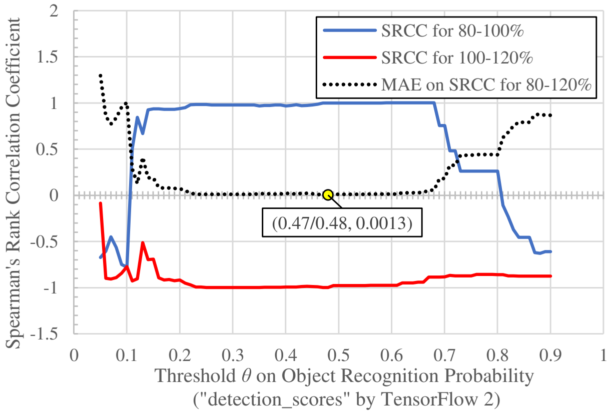


Fig. 7. The dependency of SRCC (Spearman's Rank Correlation Coefficient) between the proposed R2-B2 score (if by instance-level object recognition) and zoom [%] on its threshold $\theta$.

[13] M. Carnec, P. Le Callet, and D. Barba, "Full Reference and Reduced Reference Metrics for Image Quality Assessment," Proceedings of the Seventh International Symposium on Signal Processing and Its Applications (ISSPA'03), vol. 1, pp. 477–480, July 2003.

[14] M. Omari, A. A. Abdelouahed, M. E. Hassouni, and H. Cherif, "Improving Reduced Reference Image Quality Assessment Methods By Using Color Information," *International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)*, vol. 10, pp. 183–196, MIR Labs, 2018.

[15] Z. Wang, and A. C. Bovik, "Reduced- and No-Reference Image Quality Assessment," ıIEEE Signal Processing Magazine, vol. 28, no. 6, pp. 29–40, November 2011.

[16] V. Kamble, and K. M. Bhurchandi, "No-Reference Image Quality Assessment Algorithms: A Survey," *Optik*, vol. 126, no. 11–12, pp. 1090–1097, Elsevier, June 2015.

[17] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani "No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency," Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV'22), pp. 3989–3999, February 2022.

[18] H. Cong, L. Fu, R. Zhang, Y. Zhang, H. Wang, J. He, J. Gao, "Image Quality Assessment With Gradient Siamese Network," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 1201–1210, June 2022.

[19] Wikipedia, "Image Quality," https://en.wikipedia.org/wiki/Image_quality, July 2022.

[20] A. J. Ahumada, Jr., "Computational Image Quality Metrics: A Review," Proceedings of the SID International Symposium Digest of Technical Papers (SID'93), vol. 24, pp. 305–308, May 1993.

[21] N. Burningham, Z. Pizlo, and J. P. Allebach, "Image Quality Metrics," *Encyclopedia of Imaging Science and Technology*, pp. 598–616, January 2002.

[22] A. Horé, and D. Ziou, "Image Quality Metrics: PSNR vs. SSIM," Proceedings of the 20th International Conference on Pattern Recognition (ICPR'10), pp. 2366–2369, October 2010.

[23] S. Hattori, "Peculiar Image Retrieval by Cross-Language Web-extracted Appearance Descriptions," *International Journal of Computer Information Systems and Industrial Management (IJCISIM)*, vol.4, pp. 486–495, MIR Labs, 2012.

[24] S. Hattori, "Hyponymy-Based Peculiar Image Retrieval," *International Journal of Computer Information Systems and Industrial Management (IJCISIM)*, vol.5, pp. 79–88, MIR Labs, 2013.

[25] K. Takahashi, A. Kubota, and T. Naemura, "Focus Measurement and All in-Focus Image Synthesis for Light Field Rendering," The IEICE transactions on information and systems, Pt. 2, vol. J88-D-II, no. 3, pp. 573–584, March 2005.

[26] K. Aiba, and S. Hattori, "n/a (only in Japanese)," Proceedings of the 14th Forum on Data Engineering and Information Management (DEIM'22), E33-4 (day3 p25), https://proceedings-of-deim.github.io/DEIM2022/papers/E33-4.pdf, March 2022.

[27] W. Sun, and T. Wu, "Learning Layout and Style Reconfigurable GANs for Controllable Image Synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, arXiv:2003.11571v2, March 2021.

[28] T. Sylvain, P. Zhang, Y. Bengio, R. D. Hjelm, and S. Sharma, "Object-Centric Image Generation from Layouts," Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'21), vol. 35, no. 3, pp. 2647–2655, May 2021.

[29] S. He, W. Liao, M. Y. Yang, Y. Yang, Y.-Z. Song, B. Rosenhahn, T. Xiang, "Context-Aware Layout to Image Generation With Enhanced Object Appearance," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21), pp. 15049–15058, June 2021.

[30] L. Vladimirov, "TensorFlow 2 Object Detection API tutorial," https://tensorflow-object-detection-api-tutorial.readthedocs.io/en/latest/index.html, July 2022.

[31] V. Birodkar, "TensorFlow 2 Detection Model Zoo," https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf2_detection_zoo.md, July 2022.

[32] U. Sara, M. Akter, and M. S. Uddin, "Image Quality Assessment through FSIM, SSIM, MSE and PSNR — A Comparative Study," *Journal of Computer and Communication*, vol. 7, no. 3, pp. 8–18, March 2019.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[34] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708 December 2012.

[35] OpenCV.org, "cv::quality Namespace Reference," https://docs.opencv.org/4.x/da/d0b/namespacecv_1_1quality.html, July 2022.

[36] S. A. K. Tareen, Z. Saleem, "A Comparative Analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK," Proceedings of the 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET'08), March 2018.

[37] M. Kouchi, and M. Mochimaru, "Japanese 3-D Body Shape and Dimensions Data 2003," National Institute of Advanced Industrial Science and Technology, H18PRO-503, 2006.

[38] Amazon.co.jp, https://www.amazon.co.jp/, 11th July 2022.

[39] Dimensions.com, "Database of Dimensioned Drawings," https://www.dimensions.com/, 11th July 2022.

[40] S. Hattori, Y. Yoshida, and M. Takahara, "Improvement of Video Game Interface with Humanized Othello AIs," *The Transactions of Human Interface Society*, vol.23, no.4, pp. 459–480, November 2021.