User Evaluation of Culture-to-Culture Image Translation with Generative Adversarial Nets

Giulia Zaino, Carmine Tommaso Recchiuto, and Antonio Sgorbissa

Abstract—The article introduces the concept of image “culturization,” i.e., defined as the process of altering the “brushstroke of cultural features” that make objects perceived as belonging to a given culture while preserving their functionalities. First, we defined a pipeline for translating objects’ images from a source to a target cultural domain based on state-of-the-art Generative Adversarial Networks. Then, we gathered data through an online questionnaire to test four hypotheses concerning the impact of images belonging to different cultural domains on Italian participants. As expected, results depend on individual tastes and preferences: however, they are in line with our conjecture that some people, during the interaction with an intelligent system, might prefer to be shown images whose cultural domain has been modified to match their cultural background.

Index Terms—Cross-cultural, Image-to-image translation, Testing and Evaluation

I. INTRODUCTION

YUKIKO is an 83-year-old Japanese woman living with her son Matsuo in a traditional Japanese house. Yukiko sleeps on a futon on the tatami floor, which, she says, is very good for her back. Recently, every time Yukiko has woken up at night to go to the bathroom, she has felt a little dizzy and confused, and sometimes she has had trouble finding the light switch. Last week she fell in the dark, and nobody noticed she was lying on the floor until morning. Matsuo is worried about her safety and decided to set up a Smart Environment composed of a small table-top robot assistant named Tetsuwan Atomu, voice-activated lights, and cameras that can detect emergencies. Yukiko gladly agreed to have a camera installed in her room: “I don’t want to be a burden for my son,” she says. It’s Sunday evening: Tetsuwan greets her with a bow and then chats with her about cherry blossoms in Spring by showing pictures of beautiful trees in Kyoto on its display. “You know so many things,” Yukiko says with a smile as she lies down on her futon on the tatami floor, preparing to sleep, “Goodnight!” Tetsuwan receives an alert from one of the cameras that has recognized a person lying on the floor as an emergency, which ordinarily requires alerting Yukiko’s family. However, according to the robot’s cultural database, sleeping on a tatami floor in a traditional Japanese house is common, so this may not be an emergency. “Switch off the light, please,” says Yukiko, yawning and confirming Tetsuwan’s assessment. Tetsuwan relaxes: it would smile if it could.

The scenario above introduces a key concept: how to provide personalized interaction by taking the cultural context into account. Indeed, the way Tetsuwan Atomu greets Yukiko, the topics it chooses for conversation, the pictures it shows on its screen, and how it interprets the situation when Yukiko lies on the floor show that the assistant is culturally competent. The concept of “culture” is complex [1–5], and there is no consensus among researchers in defining it. A simple yet effective definition holds that culture is a shared representation of the world of a group of people. Then, by “culturally competent,” we mean an intelligent system that can adapt its perceptions, plans, actions, and interaction style depending on the worldview of the person it is interacting with, including their beliefs, values, language, norms, and visibly expressed forms such as customs, art, clothing, food [6–8].

Cultural factors in Robotics have been investigated in the last decade [9, 10]. Most previous approaches focused on what can make the system more or less acceptable to people of different cultures, either concerning its appearance or its behavior [11–13], including verbal and non-verbal interaction [14, 15] or social distance [16, 17]. Recently, we defined a conceptual framework to make Socially Assistive Robots

G. Zaino, C. T. Recchiuto and A. Sgorbissa are with DIBRIS, University of Genova, Via Opera Pia 13, 16145, Genova, Italy.
E-mail: antonio.sgorbissa@unige.it

1The android boy popular in Japan, also known as “Astro Boy”.

2https://cordis.europa.eu/article/id/124441-the-worlds-first-culturally-sensitive-robots-for-elderly-care
(SAR) for elderly care culturally competent [7,8] starting from research in Transcultural Nursing [3] and culturally competent Health Care [13]. However, cultural competence is crucial in all domains that could benefit from Artificial Intelligence (AI), including Education, Travel, and Business [19–21].

In this general scenario, this article focuses on the visual representation of objects in different cultures [22] and the impact on people. The anthropological study of material culture teaches us that everyday objects have different designs in diverse world areas, even when they share many similarities in terms of their functionalities. This is not a surprise: objects’ design is related to the social tasks they intend to accomplish and the material enabling them to do it [23]. Consequently, within a given geographical area and culture, some visual features tend to coherently repeat with a higher frequency, even if the presence of aesthetic universal has been found [24, 25]. With a metaphor, two objects may have the same functionalities or affordances [26] and yet be given “brushstrokes of cultural features” that make them uniquely recognizable.

Suppose now a robot interacting with a person, i.e., showing instructions on how to perform a given task such as shaving oneself or cleaning the floor: this capability may play a key role in SARs, e.g., when interacting with older people with mild forms of dementia. In the case of floor cleaning, instructions are likely to go like this: (1) Take a bucket; (2) Take a mop; (3) Take a cleaning detergent; (4) Sweep or vacuum First; (5) Fill the bucket; (6) Dip the mop; (7) Wring the mop; (8) Begin mopping (steps 6, 7, and 8 in Figure 1).

How important is it that, during interaction with people, robots and other intelligent systems show images with which the person is familiar? The advertising industry knows all too well that a product may need culturally appropriate advertisements to hit the market in diverse world areas. An ad for a detergent in Japan is unlikely to show an Italian family in an Italian house unless a specific message in this sense is required, and the opposite is true in Italy. This concept is also the basis of cross-cultural design [27, 28], a process that requires designers to understand other cultures, select cultural elements, rethink and review them, and then integrate these into a new product to satisfy target users and incorporate them into product design.

Given these premises, the article’s contribution is twofold.

- The article addresses the problem of “culturizing” images and explores Generative Adversarial Networks (GAN) [29, 35] for this purpose. With the term “culturization,” we intuitively mean altering the “brushstroke of cultural features” that make objects perceived as belonging to a given culture while preserving the perceived functionalities of objects. Specifically, we translate selected objects in the image from a source to a target cultural domain.
- The article explores the research question: “to what extent do people like images that were “culturized” to make them coherent with their culture?” To this end, we prepared an online questionnaire to be submitted to Italian participants. Then, we evaluated through statistical analysis how people perceive the cultural context of objects, their preferences for culturized and non-culturized objects, and how realistic they perceive environments patched with culturized objects.

Please notice that this work’s novelty does not consist in proposing a new network for image-to-image translation but in introducing the culturization concept, choosing a state-of-the-art GAN and a pipeline for image culturization, and defining an experimental protocol for the subjective evaluation of culturized images with recruited participants. We do not perform a comparison with previous methods because these concepts have not received much attention so far. To the best of our knowledge, the only work that shares similar ideas is [36], which uses deep learning techniques for cross-cultural design. The tool proposed requires a designer to upload a design sketch and a set of cultural images used as a reference for the style; then, it generates a culture-specific image with the same content as the uploaded sketch and the cultural style of the selected style image. However, the focus is different from our work. First, the article explores the designers’ perception of using the tool for cross-cultural design, but not the subjective perception of end-users that will ultimately use the designed object. Second, the style transfer module used in [36] can only alter the texture of images [37]. In contrast, we search for solutions capable of changing all visual design elements, including color, material, pattern, and form.

Section II describes the state-of-the-art related to image-to-image translation using GANs. Section III presents the process to culturize objects and environments, and to evaluate people’s preferences. Sections IV and V present and discuss results. Conclusions are in Section VI.

II. STATE OF THE ART

Image culturization can be described as an image-to-image translation problem. Suppose having:

- a set of images related to each other by a shared characteristic that defines them as belonging to the same domain, referred to as the source domain S;
- a set of images, different from the first one, which defines a target domain T.

The image-to-image translation problem is the problem of learning a mapping $G : S \rightarrow T$ such that the distribution of images from $G(S)$ is indistinguishable from the distribution $T$. Otherwise said, the objective is to modify images belonging to $S$ to make them “similar” to those belonging to $T$.

Image-to-image translation [38] may benefit from GAN-based approaches. GANs are not the only feasible approach to this problem [39]; however, we do not aim to find the most performing solution for image culturization. Instead, our objective is to find a feasible solution that meets all constraints, to be later tested with human participants to confirm the relevance of this new concept. According to this rationale, and considering that GAN-based methods are the vast majority of solutions, we will limit our analysis to GANs.

A. Generative Adversarial Networks

GANs have been successful in several image-to-image translation domains, ranging from the creation of purely synthetic images (e.g., faces of people that do not exist in the real world [29, 40]) to the modification of selected features in a pre-existing image (e.g., the age or facial attributes of a
solutions for image-to-image translation [30–35] that have the additional information provided to the network. B. Image-to-Image translation

Image-to-image translation is the minimax game with value function parameters \((\theta_d, \theta_g)\) that may either be from one of the samples of the data or the output of the generator and outputs the probability that \(x\) came from the data distribution \(p_{data}(x)\) rather than from the generator. After training, both \(G\) and \(D\) will reach a point at which they cannot improve because \(p_{data}(x) = p_g(x)\), and \(D(x; \theta_d) = 0.5\). Optimal parameters \((\theta_g, \theta_d)\) are obtained by playing the following two-player minimax game with value function \(V\), maximizing the cost function over \(\theta_d\) and minimizing it over \(\theta_g\).

\[
\min_{\theta_g} \max_{\theta_d} V(\theta_d, \theta_g) = \frac{1}{2} \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \frac{1}{2} \mathbb{E}_{z \sim p(z)} [1 - \log(D(G(z)))] \tag{1}
\]

Starting from this idea, Deep Convolutional Generative Adversarial Networks (DCGANs) were developed to exploit the success of Convolutional Neural Networks (CNNs): the generator still takes in input a latent noise variable \(z\) but now generates an image through convolutional decoding operations. Unfortunately, similar to the original GANs, DCGANs do not consent to specify additional constraints on the generated data samples: a relationship between the noise \(z\) fed to the generator and the generated images exists (e.g., the network produces different faces) but we cannot control the outcome (e.g., we cannot choose to produce younger or older people’s faces). Conditional GANs (CGAN) [48] overcome this limitation by providing additional information to the generator and the discriminator during training to synthesize a fake sample with desired characteristics. The generator learns to produce realistic samples (e.g., faces) corresponding to a specific label (e.g., younger or older), whereas the discriminator learns to distinguish fake sample-label pairs from real sample-label pairs. The value function \(V\) is the same as Eq. 1 by substituting \(G(z)\) with \(G(z|y)\) and \(D(x)\) with \(D(x|y)\), where \(y\) is the additional information provided to the network.

B. Image-to-Image translation

In the following, we will review a subset of GAN-based solutions for image-to-image translation [30–35] that have interesting properties for image culturization. The analysis is limited to a small subset of the available solutions [39] since this article’s main objective is to evaluate the impact of image culturization on people rather than comparing the performance of different GANs. However, given the general idea, we might consider other solutions in the future.

All approaches typically require two datasets, each dataset containing images with homogeneous content, style, and resolution. However, some methods have the additional requirements that the images need to be paired during training: for every image of the source domain \(S\), there should be a corresponding image of the target domain \(T\). An example in this class is Pix2Pix [30]. The model postulates a transformation that modifies the source image to make it ideally belong to the target domain while preserving some of its characteristics: the goal of Pix2Pix is to learn this transformation and perform it. During training, Pix2Pix adopts a CGAN approach where the input image conditions the generator; the discriminator takes as input a generated or a real image belonging to the target domain and guesses whether the image is real or not. The network is trained with the CGAN loss function by adding a weighted term that measures the difference between the generated and real images in the pair. The main limitation is that Pix2Pix and similar approaches require large paired datasets to learn the mapping between \(S\) and \(T\): paired datasets are complex and expensive to build, especially if one aims to culturize objects into several different cultures.

In order to overcome the constraint that datasets need to be paired during training, some approaches propose a new concept, referred to as consistency of the cycle: the image \(x_S\), after the first translation \(G(x_S)\) from \(S\) to \(T\), is fed to a second generator that performs a backward translation \(F(G(x_S))\) to remap the image to the starting domain \(S\) and compares it with the original image. This concept was first introduced by CycleGAN [31], trained by adding a weighed, “cycle-consistency” L1 loss measuring the difference between \(x_S\) and \(F(G(x_S))\). CycleGAN has been used to re-create the style of impressionist painters starting from a photograph. This concept fits well with our idea: in analogy with painters representing reality with their painting style, we want to capture how a human self-identifying with a given culture would represent a particular object through “cultural brushstrokes”. CycleGAN and similar approaches [32, 49] allow for building datasets quickly by searching for images on the web (e.g., Greek and Japanese vases). However, this solution also has negative sides. First, it cannot culturize an individual object in a complex scene and keep the background unaltered. Second, the image-to-image translation works better when changes occur in the texture (from a horse to a zebra, from an orange to an apple) than when they occur in shape: CycleGAN can alter the shape of an object, but this creates artifacts and less realistic images.

It would be helpful if the network could learn which parts of the scene to translate from \(S\) to \(T\) by isolating a region of interest in the image. The solution proposed in attention-guided GAN [33] achieves this: based on the structure of CycleGAN and the concept of the consistency of the cycle, two attention networks are added, one for the generator \(G: S \rightarrow T\) and the other for the generator \(F: T \rightarrow S\), which learn to extract “attention maps.” Attention maps are trained in
parallel with generators, and provide semantic information by segmenting images into regions that must or must not be translated from $S$ to $T$. This is achieved by assigning each pixel of the map a continuous value in the interval $[0:1]$: CycleGAN is equivalent to an attention-guided GAN where the attention maps are equal to 1 everywhere. This approach and similar ones [50, 51] allow modifying individual elements of a scene without further constraints on the datasets. However, attention maps imply a higher computational load and training time. Additionally, attention-guided approaches have limitations when image-to-image translation involves changes in the objects’ shape (typically required in image culturization, see the vases on the first row of Figure 4) since they aim to keep the image background unaltered.

Solutions exist to deal with datasets that exhibit more evident changes in the shape of objects or include multiple instances of objects. To this end, InstaGAN [34] requires information on the instances to be modified, obtained through segmentation masks a priori provided: $S$ and $T$ must contain both images and masks for all elements of interest. A set of generators supplements the original generator: each generator operates on a segmentation mask and produces a translated mask in the target domain. The cross-entropy used in CGAN is modified by adding weighted terms that include the cycle-consistency loss and a background-preserving loss. InstaGAN produces high-quality images where only selected regions are modified. However, it adds a new constraint: it requires a segmentation mask associated with each object. Once again, this is problematic for image culturization based on custom-built datasets composed of images downloaded from the web.

Approaches exist that emphasize the dichotomy between high-level semantics (e.g., depending on the object’s functionalities or affordances) and low-level features (e.g., “brushstrokes of cultural features”). Unsupervised Image-to-image Translation (UNIT) [35] is a representative of this class based on CoGAN [52]. CoGAN consists of two GANs: each GAN takes a random vector as input without being conditioned to any input image. Generators are trained to produce pairs of images in the two domains $S$ and $T$; discriminators are trained to distinguish generated images from real ones. Very importantly, the weights of the first few layers of the generators and those of the last few layers of the discriminators (responsible for decoding and encoding high-level semantics) are shared. Thanks to this, CoGAN generates pairs of related images in $S$ and $T$ with the same high-level features but different low-level ones. However, CoGAN is unsuitable for image-to-image translation because it is fed with random input $z$. UNIT makes the network conditioned by adding variational autoencoders [53], trained to encode images from the two domains $S$ and $T$ into a shared latent code $z$, which is then fed to a CoGAN-like network to produce images in $S$ and $T$. This process implies a cycle-consistency constraint comparable with CycleGAN’s.

C. Choosing a solution for GAN-based culturization

GANs evaluation [54] is a challenging task. GANs are trained to reach an equilibrium situation where the generator can “cheat” the discriminator. Still, the former, taken alone, is not associated with a cost function to be minimized. Under these conditions, it is hard to predict when the generator will produce samples that fit the target probability distribution at their best. Then, researchers have defined qualitative and quantitative tools to evaluate GANs.

Qualitative assessment of the generated images is often performed using crowdsourcing platforms such as Amazon Mechanical Turk [30, 31, 33, 59]. Quantitative evaluation can be performed using such metrics as the Inception score (IS) [55] and the Fréchet Inception Distance (FID) [56] – or other metrics [59]. IS evaluates generated images based on how Inception v3, a widely-used image recognition model that attains greater than 78.1% accuracy on the ImageNet [57], classifies them. IS also evaluates if there is sufficient diversity among the generated samples, addressing the so-called problem of the “collapse of the model”: the generator learns to generate a specific image of the target domain and keeps on generating the same image (even when varying the input) because the latter is very good in deceiving the discriminator. FID also uses a pre-trained Inception v3 model to measure the quality of the generated images. Still, it does so differently, i.e., by evaluating the accuracy of the Inception v3 model in classifying GAN-generated images compared with real ones. Consequently, FID captures the similarity of generated images to real ones better than IS.

In principle, we might compare the different solutions for image culturization discussed in the previous section using FID as done in [41], which shows how FID values can significantly vary depending on specific image-to-image translation task. However, we found it more appropriate to compare different GANs on a qualitative basis to assess the extent to which each solution can meet the requirements for image culturization (FID will play a role in the next section in tuning hyperparameters of the selected GAN solution). Specifically, we observed that Pix2Pix and similar approaches require paired datasets, which is a limitation since image culturization should work on custom-built datasets, e.g., composed of images downloaded from the web. InstaGAN did not reveal a feasible solution for similar reasons since it requires a carefully designed dataset composed of images belonging to different cultural domains and the corresponding masks. Following these initial considerations, we implemented CycleGAN, AttentionGAN, and UNIT that do not pose constraints on the dataset, and then we qualitatively evaluated the results. Tests confirmed that AttentionGAN has strong limitations when an image-to-image translation requires altering the image shapes, often needed in image culturization (see the vases in the first row of Figure 4). Both CycleGAN and UNIT might be good choices: after qualitative evaluation, and since CycleGAN is the basis for many other approaches (including InstaGAN that might be reconsidered in the future), we finally selected it as the best candidate for image culturization. However, we do not see any obstacles in implementing the same procedure described in the next Sections with a different GAN.

III. MATERIALS AND METHODS

This section describes the process adopted to culturize images using CycleGAN and the study design to explore our
A. Culturization pipeline

In the following, we refer to $S$ as the set of images defining a start domain (e.g., Greek vases), $T$ as the set of images defining a target domain (e.g., Chinese vases), $S \rightarrow T$ as the images generated starting from $S$ (e.g., Greek vases changed to Chinese style), $T \rightarrow S$ as the images generated starting from $T$ (e.g., Chinese vases changed to Greek style).

The culturization process includes three phases, Figure 2: (1) the building of datasets (phase 1), training and validation (phase 2), and the production of culturized images for interaction with people (phase 3). We did not follow the usual training-validation-test pipeline since we are only interested in producing images of sufficient quality for evaluation with people, not in assessing how the CycleGAN model performs on a holdout test set. Additionally, following a preliminary investigation, we realized that the default hyperparameters in the CycleGAN implementation\footnote{Available at github.com/junyanz/pytorch-CycleGAN-and-pix2pix} are appropriate to produce images of sufficient quality (which is good because retraining the model takes tens of hours with our hardware/software configuration). Therefore, after collecting datasets describing a given object class (e.g., vases) in the source and target cultural domains $S$ and $T$ (e.g., Greek and Japanese), we split them into a training and validation set. Then, we performed training and validation a few times to tune the number of training epochs and the dimensions of the datasets based on a qualitative and quantitative evaluation of the culturized images.

In phase 1, two fully automated steps are required by the system to build datasets, managed by a script in Python.

- **Image collection.** The script requires to enter the name of an object class (e.g., vase, sofa, pillow, lamp, etc.), a cultural domain (e.g., Greek, Indian, Moroccan, Chinese, etc.), and the number of images to download. The images are automatically downloaded from Google through the Selenium Web Driver API.

- **Preprocessing of images and object segmentation.** Downloaded images are preprocessed to make them suitable for CycleGAN: first, images are resized to the required $256 \times 256$ dimensions; then, objects are segmented using a network capable of identifying the desired objects within the image. Object detection and segmentation rely on the Faster-RCNN-inception-V2 pre-trained network. To this end, the candidate objects to extract need to be chosen, and networks are trained by labeling around 3,000 images via the labelImg tool. If multiple instances of an object are present, they are all detected and segmented.

In phase 2, three steps are required to train and validate the model to map images from $S$ to $T$ and vice versa.

- **Preparation of training and validation sets.** We consider two datasets of images $S$ and $T$ corresponding to the same object class in the two cultural domains and split them into a training set ($90\%$ of the dataset, built with an equal number of images from $S$ and $T$) and a validation set ($10\%$ of the dataset).

- **Training.** The network is trained to generate images $S \rightarrow T$ starting from $S$ as well as images $T \rightarrow S$ starting from $T$. We train the network with a GPU TESLA P100 using the GPU-accelerated NVIDIA CUDA Deep Neural Network library (about 30 hours for 200 epochs).

- **Validation.** The validation set is used to generate images $S \rightarrow T$ starting from $S$ as well as images $T \rightarrow S$ starting from $T$. Only the trained generator is used, whereas the discriminator is no longer needed. The resulting images $S \rightarrow T$ and $T \rightarrow S$ undergo a qualitative analysis to visually judge their quality and the FID score is computed to confirm that the results are acceptable. If needed, the network is trained again by changing the number of training epochs and the dimensions of the dataset.

During training and validation, we did not find a significant difference in the perceived quality and FID values when changing the number of training epochs from 150 to 400 (training time increasing from 20h to 50h and more). For example, when processing images of sofas, FID values ranged from 129.18 to 122.33 (from Indian to classic European) by increasing the number of epochs; when processing images of pillows, FID values ranged from 78.60 to 83.02 (from classic European to Indian). It must also be remembered that the range of FID values depends on the application for which GANs are used: FID is typically used to compare different approaches rather than providing an absolute estimate of the quality of the process. The work in \cite{41}, which focuses on human faces, reports FID values ranging from 23.72 to 48.71 for “glass removal”, from 16.63 to 36.17 for “male to female”, and from 93.58 to 102.92 for “selfie to anime”. Since CycleGAN has limitations in modifying the shape of objects, FID values tend to be quite high in our case: however, they do not change significantly with the number of training epochs. This result is also confirmed by the qualitative evaluation of culturized images, which motivated us to keep the default value of 200 training epochs in the chosen CycleGAN implementation. The same rationale motivated us to use training datasets of 1,000 images, which allows for a good compromise between FID values, perceived quality, and variability.

In phase 3, to culturize individual objects, it is now sufficient to feed CycleGAN with an image belonging to $S$ or $T$.

- **Choice of an object image.** The image of an object...
belonging to a cultural domain $S$ or $T$ is provided.

- **Image culturization.** The image is translated from $S$ to $T$ or $T$ to $S$ using CycleGAN generator.

However, if the objects to be culturized are part of a complex environment (as in Figure 1), four steps are required.

- **Choice of an environment image.** The image of an environment belonging to a domain $S$ or $T$ is provided.
- **Object segmentation.** The image of the environment is preprocessed using Faster-RCNN-inception-V2 to extract all objects that are a candidate for culturization.
- **Object culturization.** CycleGAN generator is used to map each extracted object from the domain $S$ to $T$ or $T$ to $S$.
- **Environment patching.** The original image is patched with the culturized objects. Currently, this step is not automated but manually performed with a tool.

The images produced are finally used for a more rewarding human-robot interaction: see, for example, Pepper showing cleaning instructions in Figure 3. However, culturally competent real-time interaction was not the focus of the experiments performed in this work. Due to the novelty of the culturization concept, we decided to explore our research question through an online questionnaire to measure the impact of image culturization on people, which allows for the involvement of a greater number of recruited participants.

**B. Study design**

To start exploring our research question (‘to what extent do people appreciate images that were “culturized” to make them coherent with their culture?’), we decided to do experiments with Italian participants only to simplify recruitment. We will consider other nationalities in future works. The general idea was to show participants images of different objects and environments belonging to European and non-European cultural domains and ask them some questions.

To this end, we considered four object classes (vases, sofas, pillows, and lamps) and, for each class, four sets of images produced in different ways: one set was downloaded from the Internet as belonging to the European culture and tradition ($E$); one was downloaded from the Internet as belonging to a different culture and tradition (Chinese, Indian or Moroccan, $O$ for other); one was a non-European object “culturized” using GANs ($O \rightarrow E$); one was a European object “culturized” ($E \rightarrow O$) using GANs. Specifically, we searched for evidence supporting the four hypotheses below through data acquisition and statistical analysis.

- **Hypothesis 1:** objects are correctly recognized by Italian participants as belonging to the $E$ culture. This hypothesis holds both when the object originally belonged to the $E$ culture and when it was culturized from $O$ to $E$.
- **Hypothesis 2:** European objects are generally preferred by Italian participants. This hypothesis holds both when the object originally belonged to the $E$ culture and when it was culturized from $O$ to $E$.
- **Hypothesis 3:** Italian participants who prefer $E$ objects tend also to prefer objects culturized from $O$ to $E$.
- **Hypothesis 4:** Participants perceive environments as more realistic when objects in the image were culturized with GANs than when we patched them with unmodified objects downloaded from the Internet.

It is worth spending a word concerning Hypothesis 2. Each person has their preferences: hypothesizing that Italian people will prefer European objects incurs the risk of stereotyping and, generally speaking, is not true. Different persons may be more or less attracted by objects belonging to their own or other cultures, which also depends on individual objects. For example, the same person may adore colorful Moroccan lamps and, at the same time, be particularly attracted by the design of a lamp produced in Sweden or France. We are aware of this and do not intend to make stereotyped claims. The objective of Hypothesis 2, together with Hypothesis 3, is to explore if there are people with whom using images that match their cultural background can be a winning strategy.

After preparing images using the procedures in section 3-1, we set up a Google Form questionnaire. The questionnaire was anonymous and included the four sections in Table 1 showing pictures and asking related questions:

1) **Personal questions:** the respondents had to declare their age (less than 20, 20–29, 30–39, 40–49, 50–59, 60–69, 70 and over) and if how often travel out of Italy (never or very rarely; yes, but mostly in Europe; yes, both in Europe and outside Europe);
2) **Recognizing culture and expressing preferences:** we showed respondents four vases, four sofas, four pillows, and four lamps. For each image, the respondents had to reply to the following two questions by assigning a score on a 5-points Likert scale.
   - Q1: “How close is this object to European culture and tradition?” (from 1: very far to 5: very close);
   - Q2: “Do you like this object?” (from 1: not at all to 5: a lot).
   For each object class (vases, sofas, pillows, and lamps), the four objects shown to the respondent belonged to different groups ($E$, $O \rightarrow E$, $E \rightarrow O$, and $O$).
3) **Comparing culturized and non-culturized objects:** we showed respondents sixteen pairwise comparisons, four of which concerned vases, four sofas, four pillows, and four lamps. For each comparison, two objects were shown side by side: the original one (which can either be $E$ or $O$) and its corresponding GAN counterpart.

\footnote{Full questionnaire available at https://bit.ly/3zjVHWL}
Fig. 4: Pairwise comparison of objects. E: downloaded from the Internet as belonging to the European culture; O: downloaded from the Internet as belonging to a different culture; O → E: non-European object culturized using GANs; E → O: European object culturized using GANs. 1st row: O vs. O → E; E vs. E → O. 2nd row: E vs. E → O; O vs. O → E; 3rd row: E → O vs. E; O → E vs. O. 4th row: O vs. O → E; E → O vs. E.

Table I: Questionnaire’s structure: after personal questions, 16 objects are shown in the section Recognizing culture and expressing preferences, questions Q1 and Q2; 16 pairwise comparisons of objects are shown in the section Comparing culturized and non-culturized objects, questions Q1 and Q2; 8 pairwise comparisons of environments are shown in the section Comparing the realism of environments, question Q1.

Fig. 5: Top: Pcult image patched with unmodified objects; Bottom: Gcult image patched with GAN-modified objects. (E → O or O → E, depending on the original one). The respondents had to reply to the following two questions. Q1: ‘Which object best represents European culture and tradition?’ (the one on the left or the one on the right); Q2: ‘Which object do you like the most?’ (the one on the left or the one on the right). We assigned a score of 1 to the winning object and 0 to the loser. The order in which original and modified objects appeared in the pair varied for each comparison. Figure 5 shows some of the pairs presented to respondents.

4) Comparing the realism of environments: we showed respondents eight pairwise comparisons. For each comparison, two environments were presented to the respondent: an environment that was modified by segmenting objects, culturizing them with GANs, and then pasting them back onto the original image (Gcult); an environment modified by substituting the original objects with objects of different cultures downloaded from the Internet, pasted onto the original image (PCult). The respondent had to reply to the following question. Q1: ‘Which image looks more realistic?’ (the one on the top or the one on the bottom), Figure 5. We assigned a score of 1 to the winning environment and 0 to the loser.

We use question Q1 in the questionnaire’s sections 1 and 2 to collect data supporting Hypothesis 1, Q2 in sections 1 and 2 to collect data supporting Hypotheses 2 and 3, and Q1 in section 3 to collect data supporting Hypotheses 4.

Notice that the questionnaire showed respondents objects
whose cultural belonging was emphasized. Greek vases were easily recognizable as belonging to the European culture; non-European vases were mostly Chinese and Japanese. Figure 4 shows that, when culturizing vases through GANs, the original Chinese or Greek drawings are recognizable in their GAN-modified counterparts: a Chinese vase with greek warriors and a Greek vase with an oriental tavern can be spotted in the Figure. Similarly, we chose European sofas designed in classic style. One could object that this kind of sofa is not common in ordinary European houses; however, we wanted to avoid international, modern-style sofas that may not be immediately recognizable as European. Pillows were probably less recognizable as European / non-European since colorful pillows are customary also in ordinary European houses. Finally, the cultural belonging of lamps is evident from shape and texture: see the moon-shaped Moroccan lamp versus its European counterpart, and the European “night table” lamp rethought in a Moroccan style.

The Google Form was online from 18/09/2021 to 7/10/2021. We recruited participants through public social networks (Facebook and Twitter) and student networks of the University of Genova. The questionnaire included the following introduction in Italian: ‘The questionnaire will show you a sequence of images and a few questions. The images show different objects and environments associated with different cultures. We will ask you to rate how close the images are to your culture and express your tastes and preferences. Please look at the images and answer truthfully based on your feelings: some questions require you to respond with a score between 1 and 5, and other questions require you to choose between two images. The questionnaire should take between 10 and 20 minutes.’

IV. RESULTS

This section describes collected data and related analyses.

A. Personal questions

Overall N=392 participants filled the questionnaire, out of which: 7.9% were less than 20 years old; 21.9% were in the range 20-29; 13.8% in the range 30-39; 20.4% in the range 40-49; 20.9% in the range 50-59; 11.5% in the range 60-60; 3.6% are 70 years or older. Concerning travels, 26% of the participants declared they never travel out of Italy or very rarely; 48.5% frequently travel out of Italy, but mostly in Europe; 25.5% travel both in Europe and out of Europe.

B. Recognizing culture and expressing preferences

Table II shows the results of the second section of the questionnaire. For each question Q1 and Q2, the table reports the average score and its standard deviation obtained by objects (vase, sofa, pillow, or lamp, both individually taken and all together) belonging to different groups (E, O → E, E → O, O), computed over N respondents.

To search for evidence supporting Hypothesis 1, we considered Q1: ‘How close is this object to European culture and tradition?’ To support Hypothesis 2, we considered Q2: ‘Do you like this object?’ In both cases, we performed statistical tests to reject a set of null hypotheses in the form H0 : µ1 = µ2 in favor of the alternate hypothesis H1 : µ1 > µ2. Depending on each test, µ1 must be interpreted as the mean score µE or µE→O assigned to E or O → E objects, which we hypothesized to be higher than the mean score of other objects. To test hypotheses, we initially performed an ANOVA test with α = 0.05 by analyzing all groups together (E, O → E, E → O, O) and a Tukey HSD test to check if pairwise groups are statistically different. Next, we focused specifically on the two groups E and O → E, and performed a series of one-tail, two-sample Welch’s T-test with α = 0.05 to compare them with each other and other groups. Due to the sample size N=392, we did not need to check distribution normality.

Table II shows that, concerning Question (Q) 1 (‘How close is this object to European culture and tradition?’),

- for any object class, E objects achieved the highest average score;
- for any object class, O → E objects achieved the second-highest average score.

The ANOVA returned that we can reject the null hypothesis for all object classes. The Tukey HSD test returned that all groups’ averages, when pairwise taken, are significantly different with the only exception of the couple (E → O, O) in the vases, sofas, pillows classes.

To confirm this result, we then performed a one-tail, two-sample Welch’s T-test to compare µE (corresponding to the highest average score) with µO→E (the second-highest), and then µO→E with µO and µE→O. We can reject the null hypothesis in all tests in favor of the alternate hypothesis with...
p < 0.05 for any object class.

In summarizing, with a focus on $\mu_E$ and $\mu_{O\rightarrow E}$ in Q1,

- $\mu_E > \mu_{O\rightarrow E} > \mu_O$ for vases, pillows, and lamps,
- $\mu_E > \mu_{O\rightarrow E} > \mu_{E\rightarrow O}$ for sofas,
- $\mu_E > \mu_{O\rightarrow E} > \mu_O$ when considering all classes together,

which supports Hypothesis 1.

Concerning Q2 (‘Do you like this object?’),

- for any object class, $E$ objects achieved the highest average score;
- for vases, the second-highest average score was achieved by $O \rightarrow E$ objects; for all other classes, the second-highest average score was achieved by $O$ objects.

The ANOVA returned that we can reject the null hypothesis for all object classes. The Tukey HSD test returned that all groups’ averages, when pairwise taken, are significantly different with the only exception of the couple ($O \rightarrow E$, $E \rightarrow O$) in the sofa class, ($E \rightarrow O$, $O$ in the pillow class, ($E$, $O$) and ($O \rightarrow E$, $E \rightarrow O$) in the lamp class.

The one-tail, two-sample Welch’s T-test returned that, for vases, the difference between $\mu_E$ (highest average score) and $\mu_{O\rightarrow E}$ (second-highest) is statistically significant with $p < 0.05$, and the same holds for the difference between $\mu_{O\rightarrow E}$ and $\mu_O$ (third-highest). For sofas and pillows, the difference between $\mu_E$ (highest average score) and $\mu_O$ (second-highest) is statistically significant with $p < 0.05$, and the same holds for the difference between $\mu_O$ and $\mu_{O\rightarrow E}$. For lamps, the difference between $\mu_E$ (highest average score) and $\mu_{O\rightarrow E}$ (second-highest) is not statistically significant with $p < 0.05$, but the difference between $\mu_O$ and $\mu_{E\rightarrow O}$ is.

In summarizing, with a focus on $\mu_E$ and $\mu_{O\rightarrow E}$ and Q2,

- $\mu_E > \mu_{O\rightarrow E} > \mu_O$ for vases,
- $\mu_E > \mu_O > \mu_{O\rightarrow E}$ for sofas and pillows,
- $\mu_E = \mu_O > \mu_{E\rightarrow O}$ for lamps,
- $\mu_E > \mu_O = \mu_{O\rightarrow E}$ when considering all classes together,

which supports Hypothesis 2 only partially because, in some cases, $O$ objects are preferred to $O \rightarrow E$ ones.

We performed a stratified analysis controlling for age and travel habits. Despite minor differences, the results were similar. However, something interesting emerged when considering only respondents with age $< 30$ (N=117), Table [III].

Concerning Q1, results were mostly confirmed: statistical analyses (ANOVA, Tukey HSD, and Welch’s T-test, $p < 0.05$) showed that, for any object individually taken and for all objects taken together, it always holds $\mu_E > \mu_{O\rightarrow E} > \mu_O$ with $p < 0.05$. However, concerning Q2, there are some differences compared to considering all age ranges: statistical analysis on vases returned, as previously, $\mu_E > \mu_{O\rightarrow E} > \mu_O$ with $p < 0.05$; statistical analysis on sofas returned, in this case, $\mu_E > \mu_{O\rightarrow E} > \mu_O$ with $p < 0.05$; statistical analysis on all objects taken together returned $\mu_E > \mu_{O\rightarrow E} = \mu_O$ as in the previous case, but now the average score of $O \rightarrow E$ is slightly higher than $O$, even if the difference is not statistically significant with $p < 0.05$. Results support Hypothesis 2 to a higher degree, but still partially.

To search for evidence supporting Hypothesis 3, we considered Q2: ‘Do you like this object?’ for the whole sample with N=391. Specifically, we considered the N participants and the four groups ($E$, $O \rightarrow E$, $E \rightarrow O$, and $O$). Next, for each participant and group, we computed the average score for all objects belonging to that group: this process yielded four N-sized vectors, one per group. Next, we computed Pearson’s $\rho$ to correlate the N-sized vector of E scores with the vectors of $O \rightarrow E$, $E \rightarrow O$, and $O$ scores: please notice that $\rho$ ranges from −1 to 1 and a higher correlation $\rho_{E,O\rightarrow E}$ (respectively, $\rho_{E,E\rightarrow O}$, $\rho_{E,O}$) means that respondents giving high scores to $E$ objects tend to give high scores to $O \rightarrow E$ (respectively, $E \rightarrow O$, and $O$) objects as well. The analysis returned $\rho_{E,O\rightarrow E} = 0.48$, $\rho_{E,E\rightarrow O} = 0.16$, $\rho_{E,O} = 0.15$ supporting Hypothesis 3: respondents giving high scores to $E$ objects tend to privilege $O \rightarrow E$ objects.

### C. Comparing culturized and non-culturized objects

Table [IV] reports the scores of each object class (vase, sofa, pillow, and lamp) and group ($E$, $O \rightarrow E$, $E \rightarrow O$, $O$) in the third section of the questionnaire, averaged over N respondents. To search for evidence supporting Hypothesis 1, we considered Q1: ‘Which object best represents European culture and tradition?’ To support Hypothesis 2, we considered Q2: ‘Which object do you like the most?’ In both cases, we showed each respondent two $E$ vs. $E \rightarrow O$ and two $O$ vs. $O \rightarrow E$ comparisons for each object class and assigned the score of 1 to the winner of each comparison. After averaging over two comparisons and N respondents per object class/group, the maximum score obtained in each cell of
Comparing the realism of environments

The analysis returned as a consequence of how scores are assigned to pairwise comparisons, we checked if the average number of times that \( E \) was selected \( \mu_E > 0.5 \) with \( \alpha = 0.05 \); in the \( O \) vs. \( O \rightarrow E \) comparison, we checked if \( \mu_{O \rightarrow E} > 0.5 \).

It can be observed that, concerning Q1 (‘Which object best represents European culture and tradition?’),
- for any object class, \( E \) objects were selected, on average, more than half of the time,
- for any object class, \( O \rightarrow E \) objects were selected, on average, more than half of the time.

The one-tail, one-sample T-test returned that
- \( \mu_E > 0.5 \) with \( p < 0.05 \) for any objects individually taken and for all objects taken together,
- \( \mu_{O \rightarrow E} > 0.5 \) with \( p < 0.05 \) for any objects individually taken and for all objects taken together,

which supports Hypothesis 1.

Concerning Q2 (‘Which object do you like the most?’),
- for vases, pillows, lamps, \( E \) objects were selected, on average, more than half of the time,
- the same happened for the \( O \rightarrow E \) group, even if average scores are lower than the \( E \) group,

The one-tail, one-sample T-test returned that
- \( \mu_E > 0.5 \) with \( p < 0.05 \) for vases, pillows, lamps, as well as for all objects taken together,
- \( \mu_{O \rightarrow E} > 0.5 \) with \( p < 0.05 \) for vases and pillows, \( \mu_{O \rightarrow E} = 0.5 \) with \( p < 0.05 \) for lamps, \( \mu_{O \rightarrow E} > 0.5 \) with \( p = 0.052 \) for all objects together,

which partially supports Hypothesis 2.

We performed a stratified analysis controlling for age: results with age < 30 (N=117) are in Table V All scores totaled by the \( E \) and \( O \rightarrow E \) groups tend to increase. When analyzing Q2 for the \( O \rightarrow E \) group, it holds \( \mu_{O \rightarrow E} > 0.5 \) with \( p < 0.05 \) for vases, sofas, pillows, and for all objects taken together. However, it holds \( \mu_{O \rightarrow E} = 0.5 \) with \( p < 0.05 \) for lamps. Results almost fully support Hypothesis 2.

To search for evidence supporting Hypothesis 3, we considered Q2: ‘Which object do you like the most?’ for the whole sample with N=391. Specifically, as we previously did in section IV-B we considered the N participants and the four groups (\( E, O \rightarrow E, E \rightarrow O, \) and \( O \)). For each participant, in this case, we computed only the average score for all objects belonging to the \( E \) and \( O \rightarrow E \) groups: this process yielded two N-sized vectors, one per group. Next, it was sufficient to correlate the two N-sized vectors by computing Pearson’s \( \rho_{E,O \rightarrow E} \); since it holds \( \rho_{E,O \rightarrow E} = -1 \) and \( \rho_{E,O} = -\rho_{E,O \rightarrow E} \) as a consequence of how scores are assigned to pairwise comparisons. The analysis returned \( \rho_{E,O \rightarrow E} = 0.66 \) (and, consequently, \( \rho_{E,O} = -0.66 \)) supporting Hypothesis 3: respondents giving high scores to \( E \) objects tended to privilege \( O \rightarrow E \) objects as well.

D. Comparing the realism of environments

Table VI reports the scores of each environment (\( G_{cult} \) and \( P_{cult} \)) in the fourth section of the questionnaire, averaged over N respondents. To search for evidence supporting Hypothesis 4, we asked Q1 (‘Which image looks more realistic?’). Specifically, we showed each respondent eight \( G_{cult} \) vs. \( P_{cult} \) comparisons and assigned a score of 1 to the winner of each comparison. After averaging the scores given by each respondent in the eight comparisons and then averaging over N respondents, the sum of the \( G_{cult} \) and \( P_{cult} \) columns is 1.

Analyses were performed using a one-sample T-test: in the \( G_{cult} \) and \( P_{cult} \) environments, computed over N respondents. The sum of the \( G_{cult} \) and \( P_{cult} \) columns is one.

| N=392 | Q1 | Q1 | N=117 | Q1 | Q1 |
|-------|----|----|-------|----|----|
| av    | 0.56| 0.44| av    | 0.64| 0.36|
| std   | 0.17| 0.17| std   | 0.16| 0.16|

TABLE VI: Questionnaire’s fourth Section. Left: all age ranges, N=391. Right: age < 30, N=117. The table reports the average score and its standard deviation relative to \( G_{cult} \) and \( P_{cult} \) environments, computed over N respondents. The sum of the \( G_{cult} \) and \( P_{cult} \) columns is one.

V. Discussion

The data show that responses may vary significantly depending on the object class considered and different objects within the same class. Also, in some cases, the original, non-European lamp is preferred to the GAN-modified version; in other cases, the opposite is true. This result is particularly evident when objects are not pairwise compared: people express their preferences, notwithstanding their cultural belonging, depending on their tastes or just because an object looks better than another, thanks to the photographer’s skill. The pairwise comparisons in the second section somehow control for confounding variables by presenting two objects with very similar shapes, dimensions, perspective, lighting, and sometimes texture and altering only the cultural context. However, based on the results, we cannot exclude that choosing different objects for pairwise comparisons might produce different outcomes.

Given these considerations, we can draw some conclusions by limiting our claims to the specific set of pictures shown.

First, it is evident from responses to Q1 in the second and third questionnaire’s sections that \( E \) and \( O \rightarrow E \) objects are rated as belonging to the European culture to a higher degree than their non-European counterpart, which is in line with Hypothesis 1. Interestingly, \( O \rightarrow E \) objects achieve a lower average score than \( E \) ones: this is likely because GANs preserve some elements of the source image in the translation. \( O \rightarrow E \) objects are perceived by respondents as not perfectly matching their expectations for a European object.
Second, it is evident from the responses to Q2 in the second and third sections that $E$ objects are preferred, on average, to all other objects, which is in line with Hypothesis 2. The only cases for which this result is not confirmed are the lamp class in the second section and the sofa class in the third section. However, it is impossible to draw similar conclusions when focussing on $O \rightarrow E$ objects: $O \rightarrow E$ vases are always preferred to $O$ vases, but the results vary for other object classes. The individual preferences of respondents may explain this result: due to immigration, it is very common in Italian shops and houses to see Moroccan, Indian, and Chinese pillows and lamps. People are accustomed to these objects and perceive them as familiar, and many respondents may have a favorable bias towards colorful objects coming from all around the world. A similar explanation may hold when considering the sofa class: the classic style sofas shown in the questionnaire are not very common in ordinary Italian houses where a more sober design usually prevails. Also, it is interesting that the strongest bias towards $O \rightarrow E$ objects is observed in the vase class. Greek vases are correctly perceived as European artifacts, and, on average, respondents prefer them to Far-eastern vases. This result is in line with the explanation above. According to the authors’ experience, decorated Far-eastern vases, as shown in Figure 4, are less frequently encountered in ordinary Italian houses than Moroccan and Indian pillows and lamps. Then, when comparing Greek and Far-eastern vases, the choice might be more determined by cultural belonging than individual preferences and familiarity.

Third, in the second and third questionnaire’s section younger respondents $< 30$ tend to prefer $E$ and $O \rightarrow E$ objects: this traditionalist attitude looks coherent with the lesser experience with other cultures and the limited interest younger people may have in furniture items. Many of them have never faced the problem of furnishing a house, possibly because they still live with their parents – in Italy, the average age when people leave their parents’ house is around 30.

Fourth, we found a positive correlation between respondents giving a high score to $E$ objects and $O \rightarrow E$ objects. Even if drawing strong conclusions about the preferences of Italians is not possible (and not desirable as it might lead to stereotypes), this result confirms the possibility to identify people that prefer culturized images that match their background during interactions with robots: a person that appreciates European objects, tend to favor both $E$ and $O \rightarrow E$ objects.

Fifth, it is evident from the responses to Q1 in the fourth section that environments modified by segmenting objects, modifying them with GANs, and re-inserting them into the original image are perceived as more realistic. Some respondents motivated their choice with a written comment to complain that objects in non-GAN images have a weird perspective and they do not merge well with the background. Not surprisingly, younger people $< 30$ appear more skilled in distinguishing between the GAN-modified and non-GAN environments, with a more marked preference for the former.

VI. CONCLUSION

This article introduced the concept of image “culturization,” proposed a process for object culturization based on GANs, and posed the research question of whether people appreciate images that were culturized to make them coherent with their culture. We preliminary explored this research question by analyzing the preferences of Italian participants towards objects belonging to European and non-European cultures and the realism of culturized environments. Overall, experiments motivate our intention to proceed further along this path: even if, as expected, not all participants prefer objects belonging to their cultural background, those who prefer European over non-European objects also tend to have a positive attitude towards objects that we culturized to be perceived as European.

Our work has three main limitations. First, from a technical standpoint, the culturization process is not entirely automated since culturized objects need to be manually reinserted into the environment. However, we do not expect this to impact the hypotheses tested with experiments. Second, we should take the experimental results with a grain of salt. When preparing the questionnaires, we selected a subset of objects and environments. However, we did not validate our questionnaire to capture a hypothetic construct “positive attitude towards European object”. Then, we cannot ensure that results may be generalized to any set of objects, even if the pairwise comparisons showing the original and culturized versions of the same object may somehow control confounding variables. Third, we implicitly assumed that it is possible to cultivate an environment by modifying the objects that it contains. This conjecture needs to be revised in the future: the texture, shape, and colors of walls and other architectural features may also play a role. However, given current GAN technology, this is likely the best we can do.

Future work will include an evaluation with non-Italian participants and real-time human-robot interaction with culturized images, not yet tested due to the Covid-19 pandemic.

STATEMENTS AND DECLARATIONS

The research involves the use of an anonymous online survey. The information obtained is recorded in such a manner that the identity of the human subjects cannot readily be ascertained, directly or through identifiers linked to the subjects. The authors declare that they have no conflict of interest.

REFERENCES

[1] C. Geertz, The Interpretation of Cultures, 2nd ed. Basic Books: New York, 1973.
[2] G. Hofstede, Culture’s consequences: International differences in work-related values. Sage: Beverly Hills, CA, 1980.
[3] M. Leininger, “Leininger’s theory of nursing: Cultural care diversity and universality,” Nurs. Sci. Q., vol. 1, no. 4, pp. 152–160, 1988.
[4] A. Henare, M. Holbraad, and S. Wastell, Thinking Through Things: Theorising Artefacts Ethnographically. Routledge: London, 2007.
[5] M. Carrithers, M. Candea, K. Sykes, M. Holbraad, and S. Venkatesan, “Ontology is just another word for culture: Motion tabled at the 2008 meeting of the group for
debates in anthropological theory,” Crit. Anthropology, vol. 30, no. 2, pp. 152–200, 2010.

[6] I. Papadopoulos, Transcultural Health and Social Care: Development of Culturally Competent Practitioners. London: Churchill Livingstone, 2007.

[7] B. Bruno, N. Chong, H. Kamide, S. Kanoria, J. Lee, Y. Lim, A. Pandey, C. Papadopoulos, I. Papadopoulos, F. Pecora, A. Saffiotti, and A. Sgorbissa, “Paving the way for culturally competent robots: A position paper,” in Proc. RO-MAN’17, Lisbon, Portugal, 2017, pp. 553–560.

[8] B. Bruno, C. Recchiuto, I. Papadopoulos, A. Saffiotti, C. Kouloughlioti, R. Menicacci, F. Mastrogiovanni, R. Zaccaria, and A. Sgorbissa, “Knowledge representation for culturally competent personal robots: Requirements, design principles, implementation, and assessment,” Int. J. Soc. Robot, vol. 11, no. 3, pp. 515–538, 2019.

[9] C. Bartneck, T. Suzuki, T. Kanda, and T. Nomura, “The influence of people’s culture and prior experiences with Aibo on their attitude towards robots,” AI Soc., vol. 21, no. 1, pp. 217–230, 2007.

[10] M. Rehm, “From multicultural agents to culture-aware robots,” LNCS, vol. 8004, pp. 431–440, 2013.

[11] T. Shibata, K. Wada, Y. Ikeda, and S. Sabanovic, “Cross-cultural studies on subjective evaluation of a seal robot,” Adv. Rob., vol. 23, no. 4, pp. 443–458, 2009.

[12] P. Rau, Y. Li, and D. Li, “A cross-cultural study: Effect of robot appearance and task,” Int. J. Soc. Rob., vol. 2, no. 2, pp. 175–186, 2010.

[13] H. Lee, H. Kang, M.-G. Kim, J. Lee, and S. Kwak, “Pepper or roomba? Effective robot design type based on cultural analysis between Korean and Japanese users,” Int. J. Softw. Eng. its Appl., vol. 10, no. 8, pp. 37–46, 2016.

[14] S. Andrist, M. Ziadee, H. Boukaram, B. Mutlu, and M. Sakr, “Effects of culture on the credibility of robot speech: A comparison between english and arabic,” in Proc. HRI’15, Portland, Ore, USA, March 2-5 2015, pp. 157–164.

[15] P. Rau, Y. Li, and D. Li, “Effects of communication style and culture on ability to accept recommendations from robots,” Comput. Hum. Behav., vol. 25, no. 2, pp. 587–595, 2009.

[16] G. Eresha, M. Haring, B. Endrass, E. Andre, and M. Obaid, “Investigating the influence of culture on proxemic behaviors for humanoid robots,” in Proc. RO-MAN’13, Gyeongju, Korea, 2013, pp. 430–435.

[17] M. Joosse, R. Poppe, M. Lohse, and V. Evers, “Cultural differences in how an engagement-seeking robot should approach a group of people,” in Proc. CABS’14, Kyoto, Japan, 2014, pp. 121–130.

[18] J. Betancourt, A. Green, J. Carrillo, and O. Ananeh-Firempong II, “Defining cultural competence: A practical framework for addressing racial/ethnic disparities in health and health care,” Public Health Rep., vol. 118, no. 4, pp. 293–302, 2003.

[19] L. Abrams and J. Moio, “Critical race theory and the cultural competence dilemma in social work education,” J. Soc. Work Educ., vol. 45, no. 2, pp. 245–261, 2009.

[20] J. Buehler, A. Gere, C. Dallavis, and V. Haviland, “Normalizing the fraughtness: How emotion, race, and school context complicate cultural competence,” J. Teach Educ., vol. 60, no. 4, pp. 408–418, 2009.

[21] J. Johnson, T. Lenartowicz, and S. Apud, “Cross-cultural competence in international business: Toward a definition and a model,” J. Int. Bus. Stud., vol. 37, no. 4, pp. 525–543, 2006.

[22] C. Knappett, Thinking through material culture: An interdisciplinary perspective. Pennsylvania, USA: University of Pennsylvania Press, 2005.

[23] J. Robb, “What do things want? object design as a middle range theory of material culture,” Archeol. Pap. Am. Anthropol. Assoc., vol. 26, no. 1, pp. 166–180, 2015.

[24] P. Hekkert and H. Leder, “Product aesthetics,” in Product Experience, H. N. Schifferstein and P. Hekkert, Eds. San Diego: Elsevier, 2008, pp. 259–285.

[25] M. M. Hardonk, Cross-cultural universals of aesthetic appreciation in decorative band patterns. Anonym Publisher, 1999.

[26] J. J. Gibson, The Ecological Approach to Visual Perception. London: Routledge, 1986.

[27] C. Chai, D. Bao, L. Sun, and Y. Cao, “The relative effects of different dimensions of traditional cultural elements on customer product satisfaction,” Int. J. Ind. Ergon., vol. 48, pp. 77–88, 2015.

[28] C.-H. Hsu, C.-L. Lin, and R. Lin, “A study of framework and process development for cultural product design,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6775 LNCS, pp. 55–64, 2011.

[29] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Networks,” in Proc. NIPS’14, Montreal, Canada, 2014, p. 2672–2680.

[30] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. CVPR’17, Honolulu, HI, USA, 2017.

[31] J. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. ICCV’17, Venice, Italy, 2017, pp. 2242–2251.

[32] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, “Learning to discover cross-domain relations with generative adversarial networks,” in Proc. ICML’17, vol. 70, Sydney, Australia, 2017.

[33] Y. Mejjaoli, C. Richardt, D. Cosker, J. Tompkin, and K. Kim, “Unsupervised attention-guided image-to-image translation,” in Proc. NIPS’18, Montréal, Canada, 2018, pp. 3693–3703.

[34] S. Mo, M. Cho, and J. Shin, “InstaGAN: Instance-aware image-to-image translation,” in Proc. ICLR’19, New Orleans, LA, USA, 2019.

[35] M. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” in Proc. NIPS’17, Long Beach, CA, USA, 2017, pp. 700–708.

[36] L. Zhou, X. Sun, G. Mu, J. Wu, J. Zhou, Q. Wu,
Y. Zhang, Y. Xi, N. Gunes, and S. Song, “A tool to facilitate the cross-cultural design and process learning,” *IEEE Trans Hum Mach Syst*, vol. 52, no. 3, pp. 445–457, 2022.

[37] L. A. Gatys, A. S. Ecker, and M. Bethge, “Texture synthesis using convolutional neural networks,” in *Proc. NIPS’15*. Cambridge, MA, USA: MIT Press, 2015, p. 262–270.

[38] S. Kaji and S. Kida, “Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging,” *Radiol. Phys. and Technol.*, vol. 12, no. 3, pp. 235–248, 2019.

[39] Y. Pang, J. Lin, T. Qin, and Z. Chen, “Image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging,” *Radiol. Phys. and Technol.*, vol. 12, no. 3, pp. 235–248, 2019.

[40] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in *Proc. ICLR’16*. San Juan, Puerto Rico, 2016.

[41] Y. Zhao, R. Wu, and H. Dong, “Unpaired image-to-image translation using adversarial consistency loss,” in *Proc. ECCV 2020*. Glasgow, UK, 2020, pp. 800 – 815.

[42] Z. Huang, S. Chen, J. Zhang, and H. Shan, “PFA-GAN: Progressive face aging with generative adversarial network,” *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 2031–2045, 2021.

[43] D. S. Tan, J. H. Soeseno, and K.-L. Hua, “Controllable and identity-aware facial attribute transformation,” *IEEE Trans. Cybern.*, pp. 1–12, 2021.

[44] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proc. CVPR’16*. Las Vegas, NV, USA, June 27-30 2016, pp. 3213–3223.

[45] M. Eitz, J. Hays, and M. Alexa, “How do humans sketch objects?” *ACM Trans. Graph.*, vol. 31, pp. 1 – 10, 2012.

[46] J. Yu, X. Xu, F. Gao, S. Shi, M. Wang, D. Tao, and Q. Huang, “Toward realistic face photo–sketch synthesis via composition-aided GANs,” *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4350–4362, 2021.

[47] H. Li, G. Li, L. Lin, H. Yu, and Y. Yu, “Context-aware semantic inpainting,” *IEEE Trans. Cybern.*, vol. 49, no. 12, pp. 4398–4411, 2019.

[48] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *eprint arXiv:1411.1784*, 2014.

[49] Z. Yi, H. Zhang, P. Tan, and M. Gong, “DualGAN: Unsupervised dual learning for image-to-image translation,” in *Proc. ICCV’17*. Venice, Italy, 2017.

[50] X. Chen, C. Xu, X. Yang, and D. Tao, “Attention-GAN for object transfiguration in wild images,” *LNCS*, vol. 11206, pp. 167–184, 2018.

[51] H. Tang, H. Liu, D. Xu, P. Torr, and N. Sebe, “AttentionGAN: Unpaired image-to-image translation using attention-guided generative adversarial networks,” *IEEE Trans. Neural Netw. Learn. Syst.*, 2021.

[52] M. Liu and O. Tuzel, “Coupled generative adversarial networks,” in *Proc. NIPS’16*. Barcelona, Spain, 2016.

[53] A. B. L. Larsen, S. K. Sonderby, H. Larochelle, and O. Winther, “Autoencoding beyond pixels using a learned similarity metric,” in *Proc. ICML’16*. New York, NY, USA, 2016, p. 1558–1566.

[54] A. Borji, “Pros and cons of GAN evaluation measures,” *Comput. Vis. and Image Underst.*, vol. 179, pp. 41–65, 2019.

[55] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” *Proc. NIPS’16*, 2016.

[56] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a nash equilibrium,” in *Proc. NIPS’17*. Long Beach, CA, USA, 2017.

[57] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proc. CVPR’16*. Las Vegas, NV, USA, 2016, pp. 2818–2826.

Giulia Zaino (M.S.) received her master’s degree in Robotics Engineering at the University of Genoa, where she explored generative models for modifying the cultural domains of objects and environments. Currently, she works as a developer of embedded software at SWHARD s.r.l.

Carmine Tommaso Recchiuto (Ph.D.) is Assistant Professor at the University of Genoa, where he teaches Experimental Robotics, ROS programming, and Computer Science. His research interests include Humanoid and Social Robotics, wearable sensors, and Aerial Robotics. He has been the Head of Software Development and Integration for the CARESSES project, aimed at endowing social robots for older adults with cultural competence. He has also been the local Coordinator for the BrainHuRo project, developing Brain-Computer Interfaces for humanoid robots’ remote control. He is the author of more than 40 scientific papers published in International Journals and conference proceedings.

Antonio Sgorbissa (Ph.D.) is Associate Professor at the University of Genoa, where he teaches Real Time Operating Systems, Social Robotics, and Ambient Intelligence in EMARO+, the European Master In Advanced Robotics. He is the Coordinator of National and EU research projects, among which the H2020 project CARESSES (caressesrobot.org). Also, he is the local Coordinator of the IENE 10 project, aimed at preparing health and social care workers to work with intelligent robots. His research focuses on Autonomous Robotics Behaviour, with emphasis human-robot interaction. He is the author of about 160 articles in international databases, a member of the Organizing Committee in top ranked robotic conferences, and Associate Editor in International Journals.