electronics

MDPI

Article

Feasibility Analysis of Deep Learning-Based Reality Assessment of Human Back-View Images

Young Chan Kwon [1], Jae Won Jang [1], Hwasup Lim [2,*] and Ouk Choi [1,*]

[1] Department of Electronics Engineering, Incheon National University, Incheon 22012, Korea; yckwon@inu.ac.kr (Y.C.K.); jjwopal@inu.ac.kr (J.W.J.)

[2] Center for Imaging Media Research, Korea Institute of Science and Technology, Seoul 02792, Korea

* Correspondence: hslim@kist.re.kr (H.L.); ouk.choi@inu.ac.kr (O.C.); Tel.: +82-32-835-8866 (O.C.)

Received: 26 March 2020; Accepted: 14 April 2020; Published: 16 April 2020

Abstract: Realistic personalized avatars can play an important role in social interactions in virtual reality, increasing body ownership, presence, and dominance. A simple way to obtain the texture of an avatar is to use a single front-view image of a human and to generate the hidden back-view image. The realism of the generated image is crucial in improving the overall texture quality, and subjective image quality assessment methods can play an important role in the evaluation. The subjective methods, however, require dozens of human assessors, a controlled environment, and time. This paper proposes a deep learning-based image reality assessment method, which is fully automatic and has a short testing time of nearly a quarter second per image. We train various discriminators to predict whether an image is real or generated. The trained discriminators are then used to give a mean opinion score for the reality of an image. Through experiments on human back-view images, we show that our learning-based mean opinion scores are close to their subjective counterparts in terms of the root mean square error between them.

Keywords: 3D human modeling; texture generation; deep learning; image reality assessment

1. Introduction

Realistic personalized avatars increase body ownership, presence, and dominance in virtual environments [1,2], so they can play an important role in social interactions as well as applications like virtual dressing rooms. To acquire a personalized avatar, photogrammetry scanners consisting of dozens of DSLR cameras can be used [1]. The scanners guarantee high fidelity but are demanding in terms of cost and space. Recently, methods based on convolutional neural networks (CNNs) have been proposed to create a personalized avatar from single front-view images of humans [3,4], reducing the cost and space as well as simplifying the capturing process. In [3], the texture of an avatar is obtained by generating the hidden back-view image from an input front-view image. Not only accurate but also realistic generation of the back-view image is essential because it determines the overall texture quality of the avatar.

To measure the accuracy of a generated human back-view image, we can use objective quality measures such as peak signal-to-noise ratio (PSNR) or structural similarity index (SSIM) [5]. The objective measures compare a generated image to its corresponding real image. In the single image-based scenario [3,4], the real back-view image of a human is not acquired. Thus, evaluation of the accuracy is limited in out-of-lab use of the back-view image generator.

Because the word 'reality' is a subjective term, subjective tests [6] provide the most reliable measure of the reality of a generated image. The human assessors can be asked to rate the reality of an image and then the results can be averaged into a mean opinion score (MOS). Subjective tests require dozens of human assessors, a controlled environment, and time [6]. To reduce the requirements, Ribeiro et al. proposed to have internet workers participate in subjective quality studies [7]. However, new costs are needed to motivate internet workers to achieve high throughput and high-quality results.

In this paper, we propose a fully automatic, cost-effective method for assessing the reality of a human back-view image. We train various CNN-based discriminators to predict whether an image is real or generated. Each trained discriminator gives the probability that an image is real, and we interpret the probability as a score. Our learning-based MOS, which mimics the subjective MOS, is computed as the average of the scores produced by different discriminators.

Unlike the subjective methods, the proposed method does not require human assessors and the controlled environment because the discriminators replace the human assessors. To train the discriminators, however, a large dataset and several days of training time are required. For evaluating the reality of human back-view images, we can build a large dataset by rendering publicly available 3D human mesh models [8]. Although the training time can hardly be reduced, our short testing time can compensate for the training time if the number of test images is large. Consequently, one can benefit from our proposed method if the training dataset can be easily obtained and the number of test images is too large to conduct subjective tests.

There exist other measures such as Inception score (IS) [9] and Fréchet Inception distance (FID) [10] for evaluating different qualities of generated images. To calculate IS, a pre-trained Inception model [11] is applied to every generated image to get a conditional label distribution. IS is then computed as the mean of the KL-divergence between each conditional label distribution and the marginal label distribution. IS quantifies the diversity of generated images, not the reality of individual images. To calculate FID, the Inception model is applied to all generated images and all real images, and their outputs from the penultimate layer of the Inception model are stored. The distributions of the outputs are modeled as multi-dimensional Gaussians, and then the Fréchet distance between the output distributions of real and generated images is calculated. FID measures the reality of a set of generated images, but it can not measure the reality of individual images. Shemelkov et al. [12] proposed a method for evaluating the performance of class-conditional GANs. In their method, by training an object-classification network using GAN-generated images and testing the classifier using real test images, 'GAN-train' accuracy is computed. Likewise, by training the network using real images and testing the classifier using the generated images, 'GAN-test' accuracy is computed. The GAN-train accuracy measures the diversity and realism of the generated images, while the GAN-test accuracy measures how realistic the generated images are. Like FID, Shemelkov et al.'s method cannot be used to assess the reality of individual images.

The remainder of this paper is organized as follows. Section 2 describes two different methods for generating human back-view images. Section 3 presents our learning-based image reality assessment method. Section 4 shows experimental results on the similarity of the proposed method to the subjective MOS. Finally, Section 5 concludes the paper.

2. Generation of Human Back-View Images

Given a front-view image of a person as input, Natsume et al. [3] use a U-Net architecture [13] to predict the corresponding back-view image so that both images will cover most of the person's body. Since both images share the same camera coordinate frame, the method does not require any post-processing steps to align the coordinate frame.

For the experiments in this paper, we implemented Natsume et al.'s method, which is based on the image-to-image translation framework [14]. As the generator, we adopted the U-Net-like architecture [15] used in [4]. The input to the generator network is a three-channel color image and its corresponding single-channel mask image, as shown in Figure 1. In practice, such mask images can be obtained by using a human silhouette detector [3] or a depth camera [4]. The input and output images are 256×256 pixels in size. A difference of our generator from that of [4] is that the output is multiplied by the mask image, suppressing unwanted background haloes.

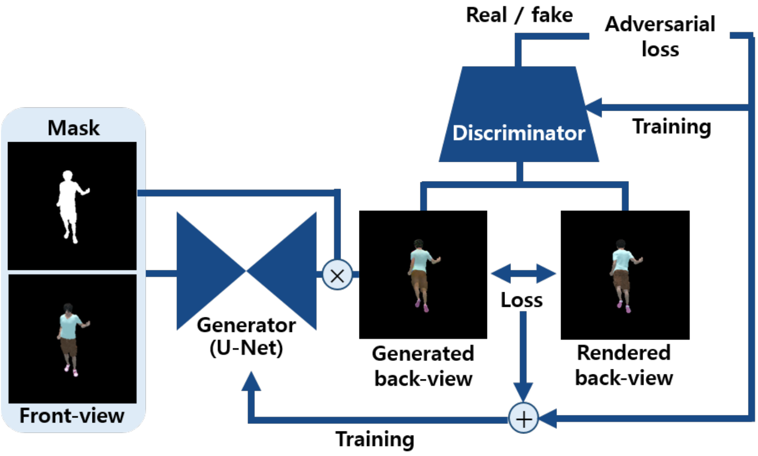

Figure 1. Generation of a human back-view image from a front-view image.

To train the generator, we need a large number of aligned human front- and back-view image pairs. In practice, however, it is difficult or impossible to obtain such aligned image pairs without reconstructing watertight 3D mesh models of the subjects. Natsume et al. used commercial datasets consisting of 3D human mesh models, which were captured by 3D scanners and then handcrafted [3]. The datasets are, however, highly expensive. In this paper, we use the publicly available Pose-Varying Human Model (PVHM) dataset [8], consisting of 10,200 mesh models of 22 different hand-crafted appearances. The model with each appearance is deformed from 200 to 1200 different poses. The models are separated into training, validation, and test sets, among which only the training set is used for training of the network parameters. The validation set is used for regularization of the network parameters, and the test set is used for the reality evaluation in Section 4. The appearances in '9200–9999' and '10,000–10,199' are used for validation and testing, respectively.

By rendering the models following the protocol in [4], we obtained a dataset of aligned human front- and back-view image pairs. The rendered back-view images are used as training targets for the generated back-view images, as depicted in Figure 1. To avoid overfitting, each mesh model was rendered four times, randomly changing the pitch angle and distance of the virtual camera, lighting conditions, the height of the model, and the color of the model's hair, clothes, and shoes. The pitch angle, distance, and height were varied from -20° to 20°, from 2 m to 3 m, and from 150 cm to 180 cm, respectively. The mean position of the light source was 5 m above the ground and its mean color was gray. We added random values to the mean position and color. Figure 2 shows sample image pairs under different rendering parameters.

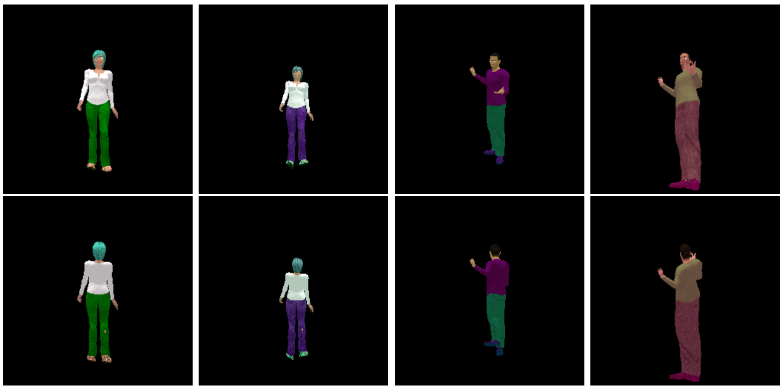

Figure 2. Image pairs obtained by rendering mesh models in the Pose-Varying Human Model (PVHM) dataset [8]. **Top**: front-view images, **bottom**: back-view images. The left and right two columns show rendered images of single models, respectively.

We use two different loss functions to train the generator. The first is the L1 loss $\mathcal{L}_{L1}$ between a generated back-view image B_G and its corresponding target back-view image B_T, defined as:

$$\mathcal{L}_{L1} = \|B_G - B_T\|_1. \quad (1)$$

Since $\mathcal{L}_{L1}$ is the sum of absolute differences in pixel values, we can expect that the trained generator will produce good objective-quality back-view images. We note that $\mathcal{L}_{L1}$ does not include an adversarial loss unlike the original Natsume et al.'s method.

The second is what was used in Natsume et al.'s method [3]. The loss function $\mathcal{L}_{SiCloPe}$ is a combination of three different losses. The first loss is a feature matching loss $\mathcal{L}_{FM}$, adopted from [16], which minimizes the discrepancy of intermediate layer activation of the discriminator between B_G and B_T. The second one is a perceptual loss $\mathcal{L}_{VGG}$, which minimizes the discrepancy of intermediate layer activation of a VGGNet-19 model pretrained for an image classification task [17] between B_G and B_T. The last one is an adversarial loss $\mathcal{L}_{adv}$, adopted from [14], which guides the generator to produce more realistic images. The total loss function $L_{SiCloPe}$ is defined as

$$\mathcal{L}_{SiCloPe} = \lambda_{FM}\mathcal{L}_{FM} + \lambda_{VGG}\mathcal{L}_{VGG} + \mathcal{L}_{adv}, \quad (2)$$

where the relative weights λ_{FM} and λ_{VGG} were set to the same values as in [3]. In the remainder of this paper, the generator obtained by minimizing $\mathcal{L}_{L1}$ and the generator obtained by minimizing $\mathcal{L}_{SiCloPe}$ are referred to as 'L1' and 'SiCloPe', respectively.

All the loss functions were minimized by the Adam optimizer with a learning rate of 2×10^4 and a mini-batch size of 4. We used early stopping for regularization of the network parameters throughout this paper, which can be approximately interpreted as L2 parameter regularization [18]. We kept monitoring the validation loss to explore the network weights minimizing the validation loss in 500 epochs. Using a computer with a single NVIDIA RTX 2080 graphics card, approximately one week was required to train each network. Figure 3 shows sample back-view images generated by L1 and SiCloPe.

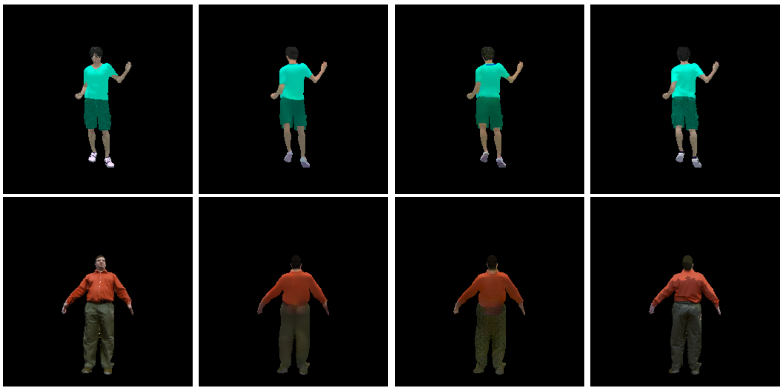

Figure 3. Samples of generated images. From left to right: input front-view images, L1-generated images, SiCloPe-generated images, and target back-view images. **Top**: PVHM dataset, **bottom**: Actual dataset.

3. Deep Learning-Based Image Reality Assessment

In subjective tests [6], human evaluators are asked to assess the quality of images. The rankings generally range from 'bad' to 'excellent' corresponding to scores from 1 to 5. The scores can be normalized to a range from 0 to 100 or from 0 to 1. If the quality is limited to reality, it might be possible to replace the human evaluators with machines by providing a number of real and generated images as a training set. The key idea of generative adversarial learning [19] is to train such a discriminator along with a generator so that the generator will produce more realistic images. Our main idea is to train multiple discriminators so that each of them will give a binary output for the reality of an image. Unlike generative adversarial learning, we do not train the discriminators along with a generator.

We can employ any discriminator as an evaluator as long as it returns reasonable outputs that are not uniform. For example, if the discriminator consistently returns either 1 or 0 irrespective of the input image, then it is highly biased and the scores should be removed before computing the mean score. In this paper, we use four different networks: the VGGNet, ResNet, InceptionNet, and DenseNet [11,17,20,21]. Each network was ranked first or second in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Our detailed architectures are VGGNet-11, ResNet-152, InceptionNet-v3, and DenseNet-161. For the networks with an input size of 224 × 224 pixels, the back-view images were resized to fit the scale. The networks have been trained on the ImageNet dataset consisting of a number of different objects. To prevent the discriminators from becoming experts in distinguishing domain-specific generated images from real ones, we only train the fully-connected layers that have been modified to provide binary class labels. This is intended to reflect the requirement of subjective testing [6] that the specialist can not be an evaluator. We used the PyTorch package [22] for the implementation, which provides pre-trained weights of the network architectures.

Each discriminator provides a binary class label along with a softmax output, which is the predicted probability of image reality. We interpret the probability as a score and calulate the mean to get a learning-based MOS. Figure 4 illustrates how the learning-based MOS is calculated.

To train each discriminator, we used the rendered target back-view images as real samples (class 1) and the L1- and SiCloPe-generated back-view images as generated samples (class 0). The generated samples are twice as many as the real samples, so we duplicated the real samples to make balanced datasets. The cross-entropy between the predicted labels and the target labels was minimized to find the parameters of each discriminator. The cross-entropy $\mathcal{L}_{CE}$ is defined for each image as

$$\mathcal{L}_{CE} = -t \log(p) - (1-t)\log(1-p), \tag{3}$$

where t is the target label, which is either 1 or 0, and p is the softmax output corresponding to class 1. At each training iteration, the loss was averaged over a minibatch and then minimized by the Adam optimizer with a learning rate of 2×10^6. We kept monitoring the validation accuracy to explore the network weights maximizing the validation accuracy in 100 epochs. The validation accuracy ranged from 71.25% to 92.84%. Using a computer with a single NVIDIA RTX 2080 graphics card, approximately one day was required to train each network. The testing time per image ranged from 39 ms to 85 ms.

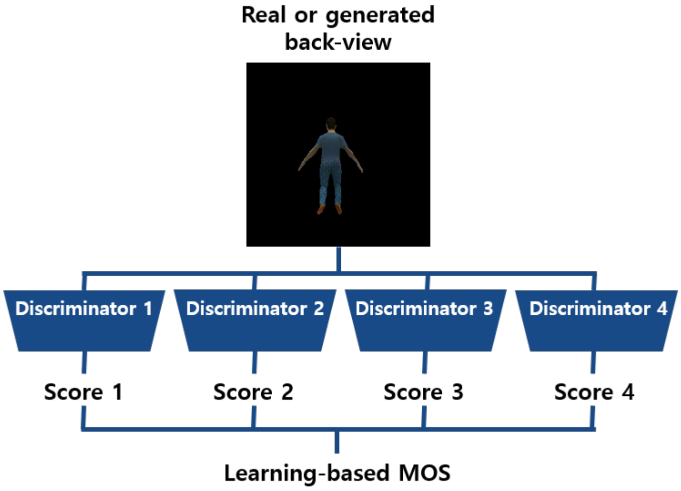

Figure 4. Computation of the proposed learning-based MOS. For a human back-view image, different discriminators compute scores to produce a mean opinion score.

Our method can be generalized to N generators and Algorithm 1 summarizes the generalized method for training a discriminator for image reality assessment.

Algorithm 1: Algorithm for training a discriminator for image reality assessment

```
Result: The parameters of the fully connected layer of discriminator $D$.
Initialization 1: Build training and validation datasets.
1. Collect the rendered back-view images of the training and validation mesh models as real
   samples.
2. Apply $N$ generators to the rendered front-view images of the training and validation mesh
   models to obtain generated samples.
3. Replicate the real samples $N - 1$ times to make balanced datasets.
4. Assign label 1 to real samples and label 0 to generated samples.
Initialization 2: Modify the fully connected layer of $D$ to output a binary label.
$Acc_{best} := 0$;
for $k = 1 : 100$ do
    Randomly shuffle training samples.
    while 1 do
        if no new minibatch is remaining then
            break;
        end
        Draw a new minibatch.
        Apply $D$ to each sample image in the minibatch to optain $p$ in Equation (3).
        Compute the average cross-entropy over the minibatch.
        Minimize the average loss by using the Adam optimizer to update the parameters.
    end
    Apply $D$ to the validation images to compute the validation accuracy $Acc$.
    if $Acc > Acc_{best}$ then
        Save the parameters.
        $Acc_{best} := Acc$;
    end
end
```

4. Experimental Results

To quantify the similarity of the proposed method to the subjective MOS test, we performed a subjective test on selected back-view images. First, we randomly selected 10 triples of rendered, L1-generated and SiCloPe-generated back-view images from the PVHM test set. Second, we randomly selected such 30 images from an 'Actual' dataset. To build the Actual dataset, we rendered 3D mesh models of real people, obtained by using the scanning method in [23]. The second row of Figure 3 shows samples of the back-view images generated from the Actual dataset. We use the Actual dataset to provide approximate results of applying the proposed method to back-view images generated from real photos. The Actual dataset enables comparisons of the generated images to their aligned realistic rendered back-view images, which can not be provided by a dataset with only real front-view images. The total number of back-view image triples of the PVHM test set and the Actual dataset is much larger, but we could not increase the number of test images due to the recommended limit of the overall duration of a subjective test session [6].

Following the ITU recommendations [6], we conducted the subjective image reality assessment test with 16 non-expert volunteers. Each image was displayed for nine seconds and a gray blank image was displayed for nine seconds between images. Prior to the main sessions, training sessions and dummy sessions of several images were used to allow the evaluators to adapt to the context.

The human assessors were asked to evaluate the reality of each image in the 1-to-5 range. Finally, we calculated the MOS for each image and normalized it to a 0-to-1 range to compare it with other measures.

We also measured our learning-based MOS of the test images. Figure 5 shows the scores produced by the four different discriminators and Figure 6 shows their mean scores co-plotted with the subjective MOS's. The leftmost 10 scores are those of the L1-generated images, the middle 10 scores of the SiCloPe-generated images, and the remainder of the rendered target images. Figure 7 shows sample test images along with their learning-based MOS's. In the top row of Figure 6, the two different MOS's are not identical but they tend to prefer SiCloPe-generated images over L1-generated images. The real images are the highest ranking in both MOS's. The results for the Actual dataset (bottom row of Figure 6) show the same trend; however, the difference between the two MOS's is slightly larger.

On the other hand, by comparing Figure 6 to Figure 5, we can see that a single discriminator's score can hardly approximate the subjective MOS. In Figure 5, individual scores are different from each other for the same image even though the discriminators have been fine-tuned on the same dataset. This phenomenon is desirable because using different discriminators was intended to mimic different opinions from different human assessors.

We calculated the root mean square error (RMSE) between the normalized subjective MOS and our learning-based MOS to quantify their difference. Table 1 shows the result. For comparison, we implemented an expert version of the proposed method 'Proposed (experts)' by training all layers of the discriminators. In Table 1, 'Masked SSIM' and 'Masked MS-SSIM' are the structural similarity index [5] and its multi-scale version [24] of the masked regions of the images, excluding the background. It can be seen that the non-expert version of the proposed method gives the closest result to the subjective MOS.

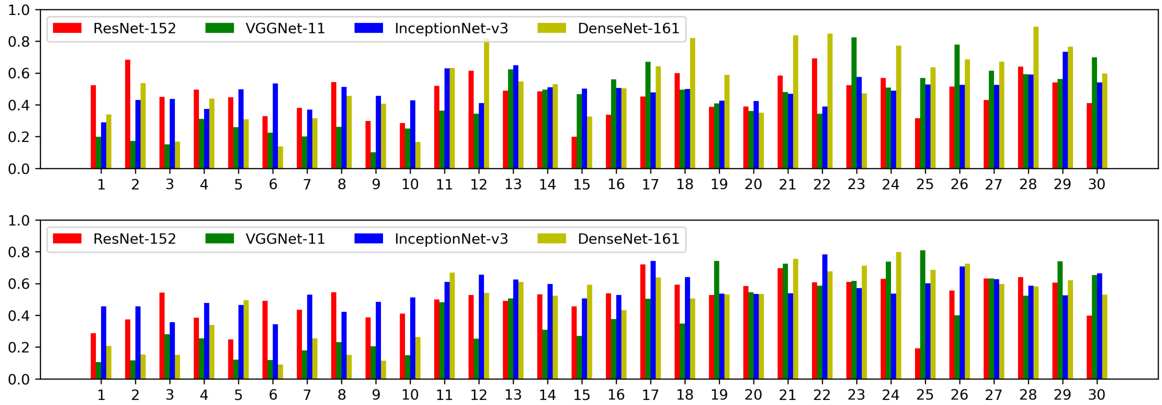

Figure 5. Scores produced by the four discriminators. **Top**: PVHM test set, **Bottom**: Actual dataset. **x-axis**: image index, **y-axis**: score.

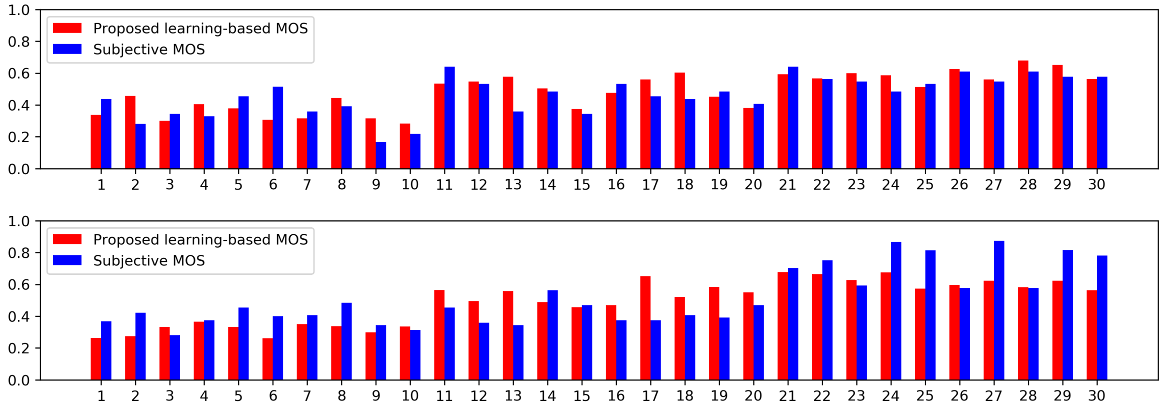

Figure 6. Comparison of the proposed learning-based MOS with the normalized subjective MOS. **Top**: PVHM test set, **Bottom**: Actual dataset. **x-axis**: image index, **y-axis**: score.

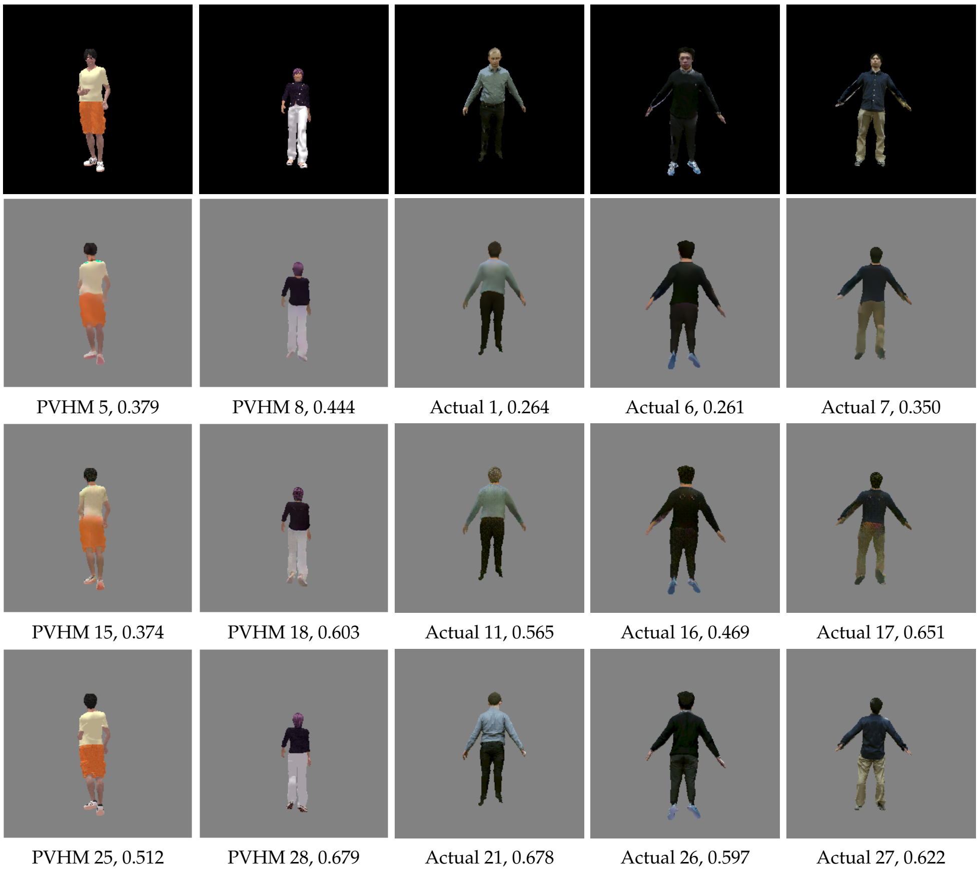

Figure 7. Sample images used for the subjective test. From top to bottom: input front-view, L1-generated, SiCloPe-generated, and target back-view images. The background of the back-view images has been filled with gray for the subjective test. Below each back-view image are the image index corresponding to that in Figures 5 and 6 and our learning-based MOS. An image can be considered as unrealistic if its score is less than 0.5, and otherwise realistic.

For further comparison, average image quality scores of the MOS test images were calculated. In Table 2, we newly calculated the average masked PSNR and FID, which could not be compared in Table 1 because PSNR is unlimited in scope and the reality of individual images cannot be measured by FID. Table 2 shows that the expert version of the proposed method gives a lower score to the L1-generated images and a higher score to the rendered images than the original non-expert version.

Because the expert version is better at discriminating real images from generated images, the scores of the L1-generated images are much lower. This is the main reason why the RMSE of the expert version is greater in Table 1.

It can be also observed that the human evaluators prefer the rendered images from the Actual dataset over those from the PVHM test set. Our post-test survey showed that the rendered images of real people contain details such as wrinkles of clothes. Although the proposed method shows the same trend, the difference is significant. Since our PVHM dataset does not contain such fine details, the discriminators did not have a chance to learn them.

According to Table 2, the objective scores, the masked PSNR, SSIM, and MS-SSIM, are consistently higher for the L1-generated images than for the SiCloPe-generated images. This result is in line with our expectation based on the characteristics of the L1 loss function. Although the objective scores are lower, the human evaluators prefer the SiCloPe-generated images. This result is consistent with the results in [25], showing the effect of generative adversarial learning.

Finally, the proposed method becomes more efficient than the subjective test as the number of test images increases. The proposed method requires several days for training; however, its test time is nearly a quarter second using a single graphics card. In contrast, the training session takes much less time than the main test session in the subjective test. However, a human evaluator takes tens of seconds to assess the quality of a single image.

Table 1. Root mean square error from the normalized subjective mean opinion score (MOS).

Method	Proposed	Proposed (Experts)	Masked SSIM	Masked MS-SSIM
PVHM	0.0928	0.178	0.377	0.467
Actual	0.138	0.170	0.214	0.348

Table 2. Average image quality scores of the selected MOS test images.

Dataset	Method	Normalized Subjective MOS	Proposed	Proposed (Experts)	Masked PSNR	Masked SSIM	Masked MS-SSIM	FID
PVHM	L1	0.349	0.354	0.184	17.6 dB	0.746	0.884	154
	SiCloPe	0.467	0.501	0.499	17.3 dB	0.707	0.872	114
	Rendered	0.569	0.593	0.737	∞	1	1	0
Actual	L1	0.384	0.315	0.136	19.3 dB	0.584	0.789	187
	SiCloPe	0.420	0.534	0.385	18.6 dB	0.514	0.752	181
	Rendered	0.736	0.620	0.689	∞	1	1	0

5. Conclusions

In this paper, we have proposed learning-based MOS, which can be used to assess image reality. In our method, various discriminators are trained to produce a binary output for the reality of an image. Their softmax outputs are interpreted as scores, and their mean is calculated to deliver the MOS. We described how to train non-expert discriminators that mimic non-expert evaluators for subjective testing. Through experiments on human texture images, we showed that the proposed learning-based MOS shows the same trend as the subjective MOS. The RMSE between the learning-based and subjective MOS's shows that the proposed method is a promising approach.

The proposed method can be more advantageous than subjective testing if the number of generators to be compared is large and the dataset of their inputs and desired outputs can be easily obtained. To obtain meaningful statistics on the performance of the generators, a large set of test images needs to be evaluated. The requirement of many human assessors, a controlled environment, and time hinders subjective testing from being a feasible solution to a large test set. Because of the relative efficiency and simplicity of testing of the proposed method, its inevitable training time can be mitigated by a large test set.

Like different discriminators giving different opinions on a single generated image, different generators produce different output images from a single input. Extending the proposed method, we will be able to build an image generating system employing multiple generators and discriminators. By training the generators along with the discriminators, the system will be able to generate different images from a single input. The generated images can be evaluated by the discriminators so that the most realistic image will be chosen as the ultimate output. As a future work, we are investigating the building of such a system.

Author Contributions: Conceptualization, H.L. and O.C.; methodology, Y.C.K. and O.C.; software, Y.C.K. and O.C.; validation, Y.C.K. and J.W.J.; formal analysis, Y.C.K. and O.C.; investigation, Y.C.K. and O.C.; data curation, Y.C.K., J.W.J., and O.C.; writing—original draft preparation, Y.C.K. and O.C.; writing—review and editing, H.L. and O.C.; visualization, Y.C.K.; supervision, O.C.; project administration, O.C.; funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported in part by the Korea Institute of Science and Technology (KIST) Institutional Program and the MSIT, Korea, under the ITRC support program (IITP-2020-2018-0-01431) supervised by the IITP.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Latoschik, M.E.; Roth, D.; Gall, D.; Achenbach, J.; Waltemate, T.; Botsch, M. The Effect of Avatar Realism in Immersive Social Virtual Realities. In Proceedings of the ACM Symposium on Virtual Reality Software and Technology, Gothenburg, Sweden, 8–10 November 2017.
2. Waltemate, T.; Gall, D.; Roth, D.; Botsch, M.; Latoschik, M.E. The Impact of Avatar Personalization and Immersion on Virtual Body Ownership, Presence, and Emotional Response. *IEEE Trans. Vis. Comput. Graph.* **2018**, *24*, 1643–1652. [CrossRef] [PubMed]
3. Natsume, R.; Saito, S.; Huang, Z.; Chen, W.; Ma, C.; Li, H.; Morishima, S. SiCloPe: Silhouette-based clothed people. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4480–4490.
4. Jang, J.W.; Kwon, Y.C.; Lim, H.; Choi, O. CNN-based denoising, completion, and prediction of whole-body human-depth images. *IEEE Access*, **2019**, *7*, 175842–175856. [CrossRef]
5. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
6. ITU-R. *Methodology for the Subjective Assessment of the Quality of Television Images*; Recommendation BT.500-14; International Telecommunication Union: Geneva, Switzerland, 2019.
7. Ribeiro, F.; Florencio, D.; Nascimento, V. Crowdsourcing subjective image quality evaluation. In Proceedings of the IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011; pp. 3158–3161.
8. Zhu, H.; Su, H.; Wang, P.; Cao, X.; Yang, R. View extrapolation of human body from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4450–4459.
9. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training GANs. In Proceedings of the Annual Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2234–2242.
10. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6626–6637.
11. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826.
12. Shmelkov, K.; Schmid, C.; Alahari, K. How good is my GAN?. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 218–234.

13. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
14. Isola, P.; Zhu, J.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976.
15. Martin Brualla, R.; Pandey, R.; Yang, S.; Pidlypenskyi, P.; Taylor, J.; Valentin, J.; Khamis, S.; Davidson, P.; Tkach, A.; Lincoln, P.; et al. LookinGood: Enhancing performance capture with real-time neural re-rendering. *ACM Trans. Graph.* **2018**, *37*, 255:1–255:14.
16. Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8798–8807.
17. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
18. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
19. Goodfellow, I.; Pouget.Abadie, J.; Mirza, M.; Xu, B.; Warde Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
21. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
22. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In Proceedings of the Annual Conference on Neural Information Processing Systems Workshop, Long Beach, CA, USA, 4–9 December 2017.
23. Lim, H.; Kang, J.; Ahn, S.C. Rapid 3D avatar creation system using a single depth camera. In Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces, Osaka, Japan, 23–27 March 2019; pp. 1329–1330.
24. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In Proceedings of the Asilomar Conference on Signals, Systems Computers, Pacific Grove, CA, USA, 9–12 November 2003; Volume 2, pp. 1398–1402.
25. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 105–114.