Photo Stand-Out: Photography with Virtual Character

Yujia Wang*
Beijing Institute of Technology
wangyujia@bit.edu.cn

Sifan Hou*
Beijing Institute of Technology
3120191002@bit.edu.cn

Bing Ning†
Beijing Institute of Fashion Technology
ningbing@bift.edu.cn

Wei Liang†
Beijing Institute of Technology
liangwei@bit.edu.cn

ABSTRACT

Extended augmented reality techniques and applications of virtual characters, like taking photography with a female warrior in Sci-Fi museum, provide a diverse and immersive experience in the real world. In different scenes, the virtual character should be posed naturally with the user, expressing an aesthetic pose, to obtain photography with the personalized posed virtual character rather than that with the immutable pre-designed pose.

In this paper, we propose a novel optimization framework to synthesize an aesthetic pose for the virtual character with respect to the presented user's pose. Our approach applies aesthetic evaluation that exploits fully connected neural networks trained on example images. The aesthetic pose of the virtual character is obtained by optimizing a cost function that guides the rotation of each body joint angles. In our experiments, we demonstrate the proposed approach can synthesize poses for virtual characters according to user pose inputs. We also conducted objective and subjective experiments of the synthesized results to validate the efficacy of our approach.

CCS CONCEPTS

• **Human-centered computing** → **Interaction design**; **Human computer interaction (HCI)**.

KEYWORDS

pose synthesis, pose aesthetic classification, pose optimization

ACM Reference Format:
Yujia Wang, Sifan Hou, Bing Ning, and Wei Liang. 2020. Photo Stand-Out: Photography with Virtual Character. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3394171.3413957

* Equal Contributors.
† Corresponding authors.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413957

Figure 1: Photographs with iconic characters: (a) photographing with a life-size statue in Christmas market; (b) photographing with museum docent in the gallery taken by Holo, a Mixed Reality application. (c) photographing with our automatically synthesized virtual character in Miaoying Temple. (Photography ©Yujia Wang, Sijing Li)

1 INTRODUCTION

When we walk in a theme park, visit a tourist attraction, attend a music festival, or enjoy a colorful parade, one of the fascinating things is to take pictures with iconic characters. For example, most people look forward to taking a fantastic photo with a life-size statue in Christmas market (Fig. 1 (a)) or with a famous star in a concert. However, most of the time, we have to experience rushed photographing and get unsatisfied photos after long queues due to limited time.

It may result in an unnatural photo since the aesthetic expression depends a lot on the relationship between the user (protagonists in the photo) posture and the virtual character posture. The techniques of computer graphics enable digital avatars and high-quality virtual characters to be synthesized in real-time, and then leveraging affordable Mixed Reality devices, e.g., smartphone, a user can pose alongside the virtual characters.

Some commercial 3D applications, e.g., Holo, develop functions of adding virtual characters into a live photo, as shown in Fig. 1 (b). Yet, the virtual characters in such applications are with fixed poses

without considering any interaction with the user, e.g., which pose the user is holding. It may result in an unnatural photo since as protagonists in a photo, the aesthetic expression depends a lot on the relationship between the user posture and the virtual character posture. However, designing poses manually for virtual characters is tedious and not scalable in practice.

In this paper, we propose a novel approach of synthesizing poses for virtual characters automatically according to the pose of the user so that they match each other from the perspective of visual aesthetic and make a realistic and vivid photo. To make such a photo, a pose should be comprised of proper angles for each body joint simultaneously so as to meet visual aesthetic criteria. In comparison to single person photography, besides the aesthetic expression of the single pose, two-person photography considers connection, interaction, and above all - feelings between two people.

Given a user pose as the input, we formulate the task of taking visually aesthetic photos for the user with a virtual character as an optimization problem. We take a similar strategy with a professional photographer to synthesize a proper pose for the virtual character, i.e. modifying and evaluating iteratively. An MCMC sampler is utilized to propose a candidate pose, then a cost function scores the proposed pose based on aesthetic criteria. The iteration continues until a proper pose is obtained. An example result of our approach is demonstrated in Fig. 1 (c).

The proposed approach can be applied to MR techniques, leading to some potential applications, e.g., making fantastic photos with virtual characters or with digital avatars. It may also help to boost the user's engagements in festivals and events.

Our main contributions in this paper are as follows:

- We introduce a novel problem of taking photos with virtual characters, whose pose synthesis is driven by the user's posture so as to output a realistic and natural photo.
- We devise a computational framework to synthesize the pose of the virtual character. Aesthetic criteria are learned and represented by a cost function, which guides an optimizer to search for a proper pose.
- We demonstrate the proposed approach for different scenes and validate its effectiveness through perceptual studies.

2 RELATED WORK

2.1 Photography Aesthetics

In modern life, photography not only represents a major technological advance, but also forms a new way of social communication [1]. Stimulated by the diversified online interactive platform, people loving photographing is growing rapidly. The core of portrait photography is to pose like a fashion model and taking professional-grade photos, which is usually contracted by amateur photographers.

Taking portrait photography needs a lot of human effort and requires the involvement of professional skills, since some aesthetic principles and photo quality assessment attributes need to be observed. Many computational approaches have been proposed to automatically analyze the aesthetics of photographic images in terms of color harmony [24], lighting [14], blur [5], photo content [14, 21], use of camera and depth [13], region composition [3, 20], etc. In addition to these basic elements, photography aesthetics can also be evaluated based on human ratings to obtain an aesthetic score distribution [12].

However, the key aim of the portrait photography is aesthetically depicting the subject (people being photographed) to convey his or her preference of the scene, i.e. the photographer seeks to convey the subjects' unique essence, feelings, and emotional complexity, which is the central goal to our modern conception of travel photography for example [8]. Just how this is done involves the use of the varied skills to show external aspects of a person, such as the pose of the subject [26].

To enhance the tour experience for amateur photographers of taking portrait photography with iconic in different scenes and inspired by the effort of designing mobile device photography assistants, such as Pose Maker [22], we explore the possibility of developing a computational approach that allows users to take photos with virtual characters in different scenes. We take the subject pose into account to synthesize such virtual characters.

2.2 Human Pose in Photography

Posing people for portrait photography is tricky, especially when you want that aesthetic look. Recently, photography assist systems have been developed to provide professional guidance to meet the visual satisfaction of photography by learning photography rules and through social content [6, 29]. There are three main kinds of systems recommend the amateurs for posing in photography: i) reference-based approach; ii) retrieve-based approach; iii) generative approach.

Reference-based Approach. Several posing references have been made available on smartphone platforms, which offers the unique possibility of directly overlaying camera view with a reference pose as visual guidance [9], such as the photo app SOVS, POING, and Holo, etc. However, the quantity of such reference poses is limited and lacks diversity, which may only contain the poses based on basic photography rules, such as leaning forward from the waist or weighting on the back leg for standing poses.

Retrieve-based Approach. The idea of guiding the human pose by retrieving existing pose from professional portrait photographs has emerged. A considerable amount of research has been done to automate pose guide, for example, Zhang *et al.* [31] proposed a framework for recommending position and poses by searching for some similar reference photos based on attention composition features. Ma *et al.* [22] proposed to retrieve an appropriate pose according to the user-provided clothing color and gender. Fu *et al.* [9] took advantage of the success of pose retrieval in previous work for portrait pose recommendation in front of a solid background.

Generative Approach. With the advancement of GANs (Generative Adversarial Networks), the problem of image generation has begun to be tackled. There is considerable work on the image generation of different human poses, such as human pose transfer [18, 23]. Balakrishnan *et al.* [2] and Li *et al.* [18] addressed the problem of synthesizing a new image of a person with the input of that person and a target 2D pose. In order to generate such image in a coherent composition, Gafni *et al.* [10] proposed a generative framework to generate an image making the person fit into the existing scene.

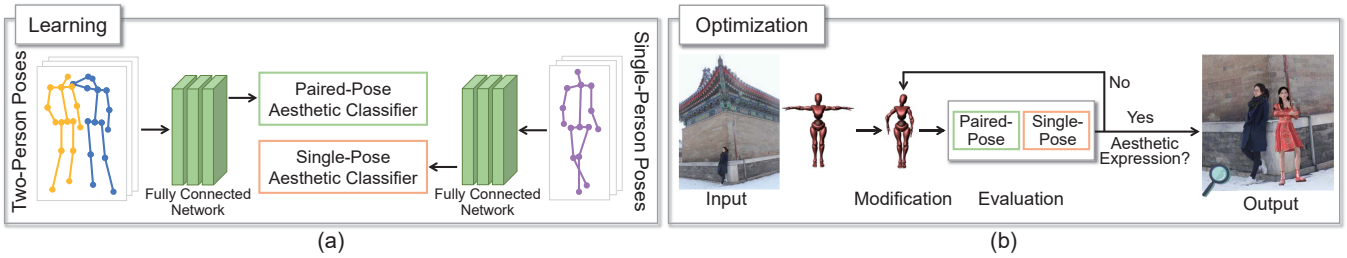

Figure 2: Overview of our approach. The framework consists of two stages, i.e. learning and optimization. The learned classifiers will be used to guide the pose synthesis optimization. (Photography ©Yujia Wang)

However, the synthesized person is limited to reference image in terms of pose diversity and could be distorted.

Previous algorithms concentrate on either single person photography or image generation with different poses, or make the image meet the requirements of photography composition. Differently, in our approach, we are concerned more about the interaction between two subjects during photography and facilitate the synthesis of the virtual character's pose according to the visual content.

3 OVERVIEW

Suppose that a user poses for being photographed, we aim at synthesizing a posed virtual character aesthetically so that the user and the virtual character are in harmony with each other in one photo. The synthesized virtual character is rendered alongside the user in the photo, which can be viewed through a mobile device, e.g., a smartphone. To achieve this goal, we devise a framework to synthesize the pose of the virtual character according to the user's pose automatically. The framework is demonstrated in Fig. 2, consisting of two stages: learning and optimization.

The purpose of the learning stage is to come up with criteria for evaluating a pose from the perspective of aesthetic expression. Because the criteria of aesthetic expression are subjective, we learn two classifiers to evaluate the aesthetic level of one pose. As shown in Fig. 2 (a), we collected two datasets: Single-Person Photograph Dataset and Two-Person Photograph Dataset. Each image in the dataset is pre-processed by extracting a corresponding skeleton for each person, represented by a pose feature vector, through an off-the-shelf pose estimation algorithm. Based on the collected dataset, a paired-pose aesthetic classifier and a single-pose aesthetic classifier are learned using fully connected neural networks, whose outputs reflect the probability of classifying the input paired-pose or single-pose as "good". The learned classifiers are utilized subsequently to guide pose synthesis optimization.

The optimization stage is demonstrated in Fig. 2 (b). Akin to the work-flow of a photographer, the optimizer modifies the virtual character's pose and evaluates the corresponding result iteratively. An MCMC sampler is utilized to explore the solution space and search for an optimal pose to match the posed user. In each iteration, the sampler proposes a candidate pose to modify, and then the pose is propagated to a total cost function to evaluate. The total cost function is defined by the aforementioned aesthetic classifiers. According to the evaluation result, the candidate is accepted or rejected, based on which the next candidate pose is proposed. The iteration continues until a proper pose is synthesized.

4 POSE AESTHETIC CLASSIFIER

It is a common practice in photography that two people cooperate with each other to make a natural and aesthetic photo. However, it is difficult to make explicit rules or define unique one-to-one correspondence for poses to make an aesthetic and harmonious photo.

For a pose with aesthetic expression in our problem, the pose should be visually aesthetic in both itself and alongside with the posed user. To this end, we learn two aesthetic classifiers to predict the aesthetic expression of poses, i.e. whether a single pose is "good" and whether two poses are "good" if they are taken by two people in one photo. The advantage is that we integrate subjective photography aesthetic evaluation into a machine learning framework through learning from a variety of professional photographs. The learned classifiers play the role of a "professional photographer", who can justify the aesthetic aspect of a pose. Thus, the classifiers can guide the optimization subsequently.

4.1 Dataset

Because of lacking a massive amount of 3D poses, we collected 2D photographs from professional photography websites instead, on which the photos are designed and selected carefully in terms of aesthetic expression.

We collected two datasets: *Single-Person Photograph Dataset* and *Two-Person Photograph Dataset*. All photos we selected are with more than 1k "likes". For *Single-Person Photograph Dataset*, each photo in the dataset contains one person. The dataset consists of 2100 photos. For *Two-Person Photograph Dataset*, each photo in the dataset comprises two people, whose poses are regarded as a pair and subsequently used in the paired-pose aesthetic classifier learning. In total, the dataset consists of 1664 photos and corresponding paired poses. Please refer to supplementary materials for the sample images of the datasets.

4.2 Learning

We utilize two neural networks to learn two aesthetic classifiers. All poses are pre-processed to extract the corresponding pose features, which are propagated to the neural network.

Pose Feature. The photos in the datasets are pre-processed to form pose features for the classifier with 3 steps. (1) The photos are detected by OpenPose [4], an off-the-shelf pose detection algorithm, to output two poses represented by the skeleton. We choose 16 keypoints, which affect pose apparently, to calculate the pose feature. The keypoints contain shoulders, elbows, torso, etc. (2) Two adjacent keypoints are connected and transformed into an angle in polar coordinates. Each pose is then represented as 15 angles, as shown in Fig. 3. (3) For single-pose, the 15 angles are used as the pose feature directly. For paired-pose in *Two-Person Photograph Dataset*, the angles of two poses are concatenated together and used as the pose feature.

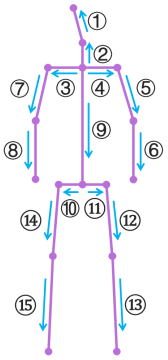

Figure 3: Pose feature.

Positive and Negative Examples. All poses in our dataset are regarded as positive examples ("Good") in the training process. It makes sense because the photos come from professional photography websites.

For negative examples ("Bad"), we synthesized 5k single-pose and paired-pose respectively by sampling pose parameters randomly. To further obtain negative examples, we recruited 3 participants, who have photography learning and practical experience, to rate the poses with aesthetic expression. The rating ranges from 1 to 5, i.e. a 1-5 Liker scale, with 1 meaning inappropriate and unsightly pose and 5 meaning the opposite. We chose 2k single-poses and 1.6k paired-poses with ratings lower than 3 as negative examples. Fig. 4 shows some examples of our training data in the form of a skeleton.

Neural Network. Both classifiers are with the same fully connected neural network structure to classify the pose aesthetic. The model is designed to have three fully connected layers (with 256, 128, and 2 output sizes). ReLU activation is applied for the first two layers and Sigmoid activation is applied for the last one. We use Binary Cross Entropy as the loss function.

In each learning task, 80% of poses were fed into the network and were split 80% for training and 20% for testing in each epoch. The network is trained with a batch size of 32 and a momentum of 0.9. The learning rate is set to be 0.0001 and weight decay is 0.1.

Evaluation. We test the performance of our models on the 20% test set of the collected *Single-Person Photography Dataset* and *Two-Person Photography Dataset* and achieve accuracies of 99.67% and 99.52% for single run evaluation, respectively. We also compared the performance of our model against different architectures, such as SVM, achieving prediction accuracies of 84.23% and 86.88%, and found that the fully connected network achieves the best results.

5 POSE SYNTHESIS

A pose for a virtual character is defined as a set of Euler angles, written as $\{\Theta_i\}$, $i \in [1, 11]$. $\Theta_i = (\theta_p, \theta_y, \theta_r)$ is one joint of the virtual character, representing pitch, yaw, and roll of the joint.

Given a user presenting a specific pose Φ_I, our approach suggests a virtual character's pose with a parameter $\Theta^* = \{\Theta_i\}$ such that they are harmonious with each other and visually aesthetic. Note that

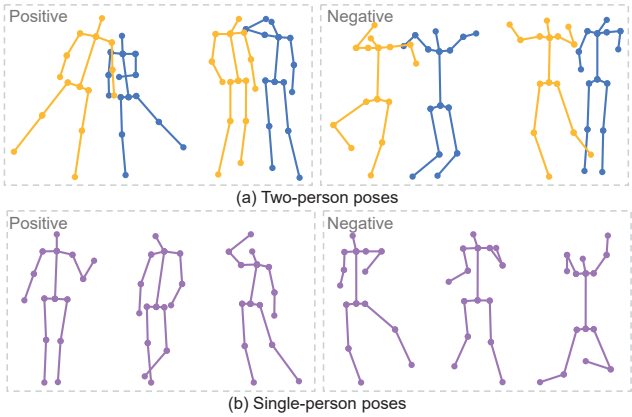

Figure 4: Example images in our training dataset, including paired-poses of two-person and single-person poses.

Φ_I is a 15D pose feature along 2D plane, extracted in the same way as the pre-process in Sec. 4.2. $\{\Theta_i\}$ is a 33D parameter in 3D space, which is used to drive a virtual character's pose. By pre-processed, the pose parameters of the virtual character can be represented as a 15D pose feature.

We solve the problem of searching for the parameter Θ^* by optimizing a cost function.

5.1 Cost Function

We define a cost function to evaluate a synthesized pose of the virtual character from two aspects: the aesthetic of the paired-pose with the user and the aesthetic of the single-pose itself.

$$C(\Theta, \Phi_I) = C_p(\Theta, \Phi_I) + \lambda C_s(\Theta) \tag{1}$$

$C_p(\Theta, \Phi_I)$ is the cost of paired-pose. It evaluates the visual aesthetic expression of putting the synthesized pose Θ and input pose Φ_I together. The virtual character with the pose of Θ is rendered to generate an image using the frontal view. Lambertian surface reflectance is assumed and the illumination is approximated by second-order spherical harmonics [25]. The rendered image is then represented by a 15D pose feature Φ_s. Both Φ_s and Φ_I are concatenated and propagated to the corresponding classifier. The cost is defined as:

$$C_p(\Theta, \Phi_I) = C_p(\Phi_s, \Phi_I) = 1 - \frac{\exp(x_1)}{\exp(x_1) + \exp(x_2)} \tag{2}$$

where $[x_1, x_2]^T$ is the output of the fully connected layer. x_1 and x_2 reflect the possibilities of Θ matches with the given user's pose or not, respectively.

$C_s(\Theta)$ represents the cost of the single-pose, which is defined similarly with the paired-pose cost. λ is a parameter to balance the weights of two cost terms. In our experiments, λ is set as 0.5 by default.

5.2 Optimization

For the cost function $C(\cdot)$ in Eq. 1, we adopt a Markov Chain Monte Carlo (MCMC) sampler to explore the pose space iteratively. In

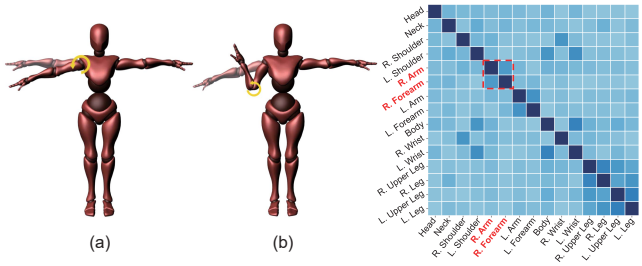

Figure 5: Left: example correlated joints and the corresponding pose changes. The rotation of the upper arm will drive the movement of the elbow (shown in (a)), which allows the flexion and extension of the forearm relative to the upper arm (shown in (b)). Right: the visualized 2D feature element distribution is shown in the heatmap pattern. Please refer to supplementary materials for the correspondences between joints and 2D features.

each iteration, the sampler proposes a move Θ', representing a candidate pose. The candidate is evaluated using the cost function. The proposed pose is accepted or rejected according to the Metropolis Hastings's acceptance probability [11]:

$$\mathcal{A} = \min\left\{1, e^{\frac{1}{T}(\mathrm{C}(\Theta,\Phi_I)-\mathrm{C}(\Theta',\Phi_I))}\right\}, \tag{3}$$

where T is the temperature of the simulated annealing process. We set the value of T empirically as 1.0 at the beginning of the optimization, allowing the optimizer to explore the solution space more aggressively. The value of T drops by 0.05 every 10 iterations of the optimization, allowing the optimizer to refine the solution near the end of the optimization.

We use two strategies to propose a 'move': correlation move and prior move.

Correlation move. When posing a person to give aesthetic expression, some joints of the pose are correlated. For example, in most cases, the rotation of the upper arm on roll axis will drive the elbow up and down (shown in Fig. 5 (a)), which allows for the flexion and extension of the forearm relative to the upper arm (shown in Fig. 5 (b)).

To explore the pose space efficiently, we learn correlations among 2D pose features. A correlation analysis over the collected photography datasets in Sec. 4.1 is applied to model the relation between every two pose features. The correlation visualized with a heatmap style is illustrated in the right column of Fig. 5. It turns out that some features have strong correlations, e.g., right upper arm and right forearm. Note that, the correlation is based on the 2D pose feature. Roughly, according to the calculation of pose feature, we can project the correlated features back to the corresponding joints in the parameter space.

If the sampler takes a correlation move, we randomly select one joint, whose correlated joints are sampled together. Please refer to supplementary materials for the detailed statistical results, the correlation among joints, and the correspondences between joints and 2D features.

[1]https://en.dpm.org.cn/

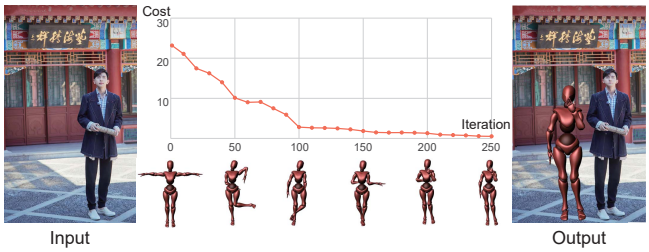

Figure 6: Pose optimization from an initialized T-pose. As the optimization process proceeds, the pose of the virtual character is iteratively updated until the two poses (i.e. the presented user's pose and the optimized virtual character's pose) converge to the desired visual aesthetic expression. The optimization finishes in about 0.25 seconds. (Photography ©The Palace Museum [1])

Prior move. According to our observation, when a person presents a pose to take an appealing photo, the joints he/she adjusts often concentrate in a subset. If the sampler explores those joints, it may have more chances to approach the solution.

Therefore, based on the two collected datasets, we calculate the change of each joint with respect to a natural standing pose. For the ith joint, the change is defined as $\Delta\theta_i = \sum_{n=1}^{N} \frac{|\theta_i - \theta_i^0|}{Z_i}$, where θ_i^0 is the value of the ith joint with a natural standing pose; Z_i is the maximum of the change for the ith joint; N is the number of the poses in the dataset. If the sampler takes a prior move, the joint is selected according to the probability of $\frac{\Delta\theta_i}{\sum_i \theta_i}$. The corresponding joints Θ_i is used to generate a new candidate as $\Theta_i' = (\theta_p + \alpha, \theta_y + \beta, \theta_r + \gamma)$, where $(\theta_p, \theta_y, \theta_r)$ is the candidate of last iteration; α, β, γ is a random value of $[0, 5]$.

We use correlation move and prior move with to probabilities α and $1 - \alpha$ respectively. In our experiments, we set $\alpha = 0.3$ by default to slightly favor selecting a prior move, corresponding to more local refinement.

The optimization is initialized by a T-pose, and continues until the absolute change in the total cost value is less than 5% over the past 25 iterations. Fig. 6 shows the iterative optimization process. The supplementary videos include animations of the optimization process.

6 EXPERIMENTS

In this section, we discuss several objective and subjective experiments conducted to evaluate the effectiveness of our pose synthesis approach. We implemented our approach using Phython3.6 and Maya2019. We ran our experiments on a PC equipped with 16GB of RAM, a Nvidia Titan X graphics card with 12GB of memory, and a 2.60GHz Intel i7-5820K processor.

In order to meet the basic rules of photography composition [15], we paralleled analyze the scene to determine the position of the virtual character. We first use off-the-shelf techniques to detect the salient region [7] and the planar region [19] of the 3D scene. Then the virtual character is placed within the planar with the user. The position is sampled according to the distribution of conventional position, estimated from overall data in the *Two-Person photograph*

Figure 7: Different results of *Random Synthesis*, our approach, and *Professional Synthesis* that used in our experiments. (Photography ©Yan Zhang, Min Gong)

Dataset, and minimizes the occlusion of the salient region. Please refer to supplementary materials for the results of our rendered virtual characters in different photographs.

6.1 Compared approaches

To verify the proposed approach, we compare with two baseline approaches of virtual character pose synthesis:

- *Random Synthesis.* The parameters of 11 joints are randomly synthesized;
- *Professional Synthesis.* We recruited three professional photographers who have been engaging in relevant works for 5 years. The professionals were asked to pose the virtual character based on the fixed camera view and the presented user, the process of which is similar to the retrieve-based pose recommendation systems [22, 31] but breaking the bottleneck of the limited reference poses.

We compared the results of these approaches in objective and subjective experiments.

Validation Dataset. The validation dataset consists of 25 scenes, such as the Forbidden City and Science Museum, in each of which there is one posed user. We synthesized 3 virtual character poses for each scene according to the user's pose using three approaches aforementioned. In order to avoid the influence of the virtual character's appearance on the rating results, we use an *X Bot* model to be the photographed virtual character avoiding appearance bias. The different posed virtual characters are placed in the same position, which is pre-processed according to the input scene and the user presented pose.

Fig. 7 depicts two sets of compared results in our experiments. Please refer to supplementary materials for all results.

Table 1: Demographics of study participants. *G=Gender (M=Male; F=Female). Occ.=Occupation* (UOR=Unemployed or retired person; STU=Students; EDU=Educators; MER=Merchants). *Phot. Exp.=Photography Experience* (NPE=People with no photography experience; ALS=Amateurs with low skills level; AHS=Amateurs with high skills level; PRO=Professional photographers).

G.	Age	Occ.	Phot. Exp.
M: 24	<20 : 6	UOR: 3	NPE: 5
F: 18	20-30: 26	STU: 30	ALS: 26
	30-40: 7	EDU: 3	AHS: 9
	40-50: 3	MER: 6	PRO: 2

6.2 Objective Evaluation

We measured the objective performance of the compared approaches with respect to pose aesthetic expression and synthesis time.

(1) Pose aesthetic expression. Due to the lack of computational methods for multi-person photo aesthetics evaluation, we use the cost function as a reference metric, i.e. for the photographs synthesized by different approaches, we compute the average aesthetic scores ($score = 1 - C(\cdot)$). A higher score (a lower cost) means the synthesized pose is regarded as "good" from the perspective of aesthetic classifiers. The advancement of the computational method of image aesthetics analysis will help such objective evaluation, or our trained classifier could serve as a baseline approach in this research field.

Among the results of three compared approaches, our approach attained the highest score ($M = 0.99, SD = 0.003$), closely followed by professional approach ($M = 0.89, SD = 0.24$). The strategy of training the pose aesthetic classifier on professional photographs has a comparable capability to professional evaluation. The random synthesis results obtain the lowest average score ($M = 0.35, SD = 0.37$), which verifies the synthesized pose without any constrains can not satisfy the aesthetic expression standards in most cases.

(2) Synthesis time. For each scene and subject, we recorded the synthesis time of the pose synthesis process for each approach. The results show that the *Random Synthesis* uses the least time to synthesize a pose ($M = 0.001$ s, $SD = 0.0001$ s). Our approach synthesizes a pose much faster ($M = 0.28$ s, $SD = 0.02$ s), including 0.02 s for position determination and 0.26 s for 250 iterations in the optimization process, compared to *Professional Synthesis* ($M = 15.63$ mins, $SD = 1.84$ mins). The experts claimed that they need to first analyze the scene composition and the presented user pose in a very short time and initialize the pose of the virtual character correspondingly. Then, they spent about 10 minutes to fine-tune the virtual character's pose to meet the aesthetic criteria. The results show that the synthesis time of our approach enables real-time applications, photograph apps on a smartphone for example.

6.3 Subjective Evaluation

Next, we carried out user studies to evaluate the effectiveness of our approach and the aesthetic experience subjectively. We recruited 42 participants, with reported normal or corrected-to-normal vision and no color-blindness. The detailed participants demographics are

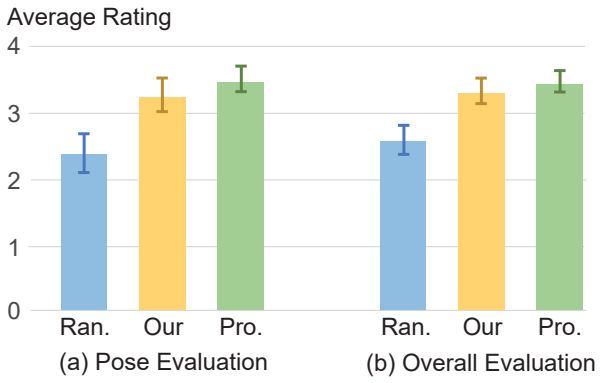

Figure 8: Average user ratings of pose aesthetic expression and overall experience of different approaches (i.e. *Random Synthesis*, our approach, and *Professional Synthesis*) in subjective evaluation.

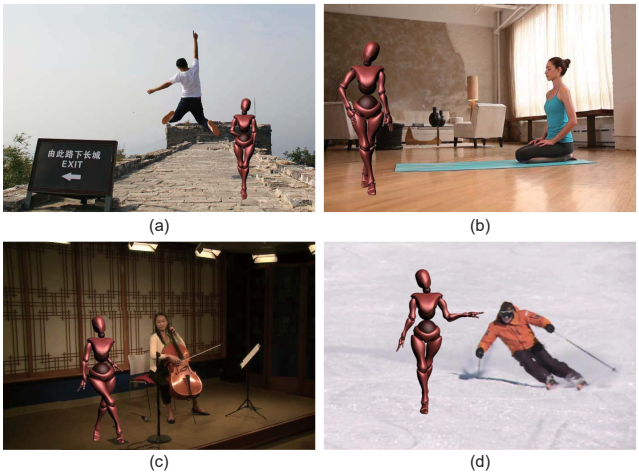

Figure 9: Results of "special" user input poses. (Photography (a) ©Chenhao Li)

shown in Table 1. As shown, the participants represent a diversity of backgrounds in terms of gender (24 men, 18 women), age (range is 18 to 50 with a mean of 24.82), occupation (from people who are unemployed or retired, to those who are students, educators, and merchants). These participants had a range of photography experience. Before each study, the participants were given a task description and encouraged to ask any questions. The participants were seated 35 cm in front of a screen (with 1440×900 resolution).

Our subjective experiment consisted of two parts: i) virtual character's pose evaluation (verifying the aesthetic expression of our synthesized pose respect to the presented user pose); ii) overall evaluation (verifying the experience enhancement of photographing with posed virtual character in different scenes). The rating ranges from 1 to 5, i.e. a 1-5 Liker scale, with 1 meaning inappropriate and unsightly pose and 5 meaning the opposite. The results were shown to the participants in a random order so as to avoid bias. We conducted a semi-structured interview about users' experience to explore the factors influencing the ratings.

Outcome and Analysis. The statistics of the average ratings across the 25 scenes on pose aesthetic expression and overall evaluation are shown in Fig. 8. The *Professional Synthesis* obtained the highest rating in both two evaluations (pose: $M = 3.51, SD = 0.39$; overall: $M = 3.48, SD = 0.32$), followed by our approach (pose: $M = 3.27, SD = 0.50$; overall: $M = 3.34, SD = 0.38$) and *Random Synthesis* (pose: $M = 2.40, SD = 0.58$; overall: $M = 2.61, SD = 0.43$). In all cases, our syntheses are more preferable than random syntheses and are comparable to professional syntheses.

To ascertain that our results are efficacious, we performed One-Way ANOVA test on the average ratings of each scene, using a significance level of 0.05. The results suggest that our approach obtained higher score than *Random Synthesis* significantly: pose ($F_{[1,49]} = 32.194, p < .05$); overall ($F_{[1,49]} = 41.088, p < .05$). On the contrary, there are no significant differences between our results and *Professional Synthesis* results: pose ($F_{[1,49]} = 3.635, p = 0.063 > .05$); overall ($F_{[1,49]} = 2.113, p = 0.153 > .05$). It turns out that our approach can synthesize satisfactory poses for the virtual character for different scenes and different presented user poses, which can enhance the photography experience and is particularly useful for novice photographers simultaneously.

To further verify that the pose expression contributes to the overall experience, we computed Bivariate (Pearson) correlation coefficients between the ratings of the overall experience and pose expression. There are positive correlations between them ($r = 0.810, p < .05$). The result suggests that improvement of the pose aesthetic expression correlates with increases in the overall experience. This supports our adopted strategy, i.e. showing virtual character by considering the pose expression.

User Feedback. Most users commented that they had better experiences of photographs with the posed virtual characters. For photographs taken in different scenes, users commented that they could feel out the virtual character changes accordingly to the user's pose. However, some participants commented that the absent face, garment, and hairstyle could affect the overall experience. This is a limitation posed by the fact that photography aesthetics perception of different users is inconsistency, thus we try to eliminate such bias by using the *X Bot* model.

Besides the considered factors for posing the virtual character, some users commented that the pose interactions with some specific users or scenes is slightly awkward, e.g., the arm pose of the virtual character is unnatural for the child user, or the pose is not coordinated with the photography props (such as an umbrella). Some users stated that the virtual character without facial expression could not fully satisfy the needs during photographing. Moreover, some users suggested that it would be interesting to make dynamic short video recording with the synthesized virtual characters applied with speech synthesis [28]. Such feedbacks give us some interesting insights about considering the personalized information of different users, e.g., gender, age, emotional states, and scene context in synthesizing the virtual character.

7 CONCLUSION

In this paper, we propose a new problem of taking photos with virtual characters considering aesthetic expression. To achieve this

goal, we devise a computational framework to automatically synthesize an aesthetic pose for a virtual character according to a given user's pose.

Our approach leads to a variety of potential applications, such as entertainment and advertisement. For example, it is popular to have virtual characters in games. Users often treat such virtual characters as their virtual friends. By our approach, users can take a visually aesthetic photo with the virtual world characters whenever they want, just like taking photos with a real friend since the virtual character can interact with users. Another potential application is for advertising. For example, when a new movie is released, it is common to place some protagonist's life-size statues to attract customers. By our approach, a virtual character can interact with customers so as enrich the advertisement form and reduce the costs simultaneously.

Limitation and Future Work. As the performance is limited to the aesthetic classifiers trained on the collected datasets, diverse user input poses experience failures in the virtual character's pose synthesis. Fig. 9 shows some results of "special" user input poses. Although our datasets are collected from professional photograph website, the pose diversity is limited by dataset quantity, for example, there are no similar training data to that in Fig. 9 (a) and (b). The advancement of training strategies, the availability of large-scale datasets for training, and the quality of *Single-Person Photograph Dataset* and *Two-Person Photograph Dataset* will help synthesize more realistic and diverse virtual character poses.

Our current approach considers the interactive pose between two subjects. As shown in Fig. 9 (c) and (d), it would be interesting to consider more context beyond the pose data in photos, such as photography scene attributes (e.g., dynamic or static), scene semantics (e.g., key objects layout [17]), users' personal attributes (e.g., gender, age, weight, and height) and preferences (e.g., the relationships to the users, like friends, couples, or rivals), key objects in the scene (e.g., photography props), etc. Moreover, group people photography is also an interesting direction for future work since photographs with multi-subjects featuring friends and families are common.

Besides our considered aesthetic performance of the synthesized pose, some factors may also affect the aesthetic expression, which is observed in our experiments. We believe it is an interesting extension to model more components using our optimization framework, e.g., virtual character's clothes [30], facial expression [16], and head pose [27]. Such factors could drive more vivid virtual characters, not limited to poses, but also yield more diverse appearance, thus achieving more attractive portrait photography.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61972038. We thank Sijing Li, Min Gong, Chenhao Li, and Yan Zhang for their help with providing photographs taken in different scenes.

REFERENCES

[1] Patricia C Albers and William R James. 1988. Travel photography: A methodological approach. *Annals of tourism research* 15, 1 (1988), 134–158.
[2] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. 2018. Synthesizing images of humans in unseen poses. In *CVPR*. 8340–8348.
[3] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. 2010. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the 18th ACM international conference on Multimedia*. 271–280.
[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*. 7291–7299.
[5] Chunhua Chen, Wen Chen, and Jeffrey A Bloom. 2011. A universal reference-free blurriness measure. In *Image Quality and System Performance VIII*, Vol. 7867. International Society for Optics and Photonics, 78670B.
[6] Bin Cheng, Bingbing Ni, Shuicheng Yan, and Qi Tian. 2010. Learning to photograph. In *Proceedings of the 18th ACM international conference on Multimedia*. 291–300.
[7] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. 2014. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 3 (2014), 569–582.
[8] Cynthia Freeland. 2007. Portraits in painting and photography. *Philosophical Studies* 135, 1 (2007), 95–109.
[9] Hongbo Fu, Xiaoguang Han, and Quoc Huy Phan. 2013. Data-driven suggestions for portrait posing. In *SIGGRAPH Asia 2013 Technical Briefs*. ACM, 29.
[10] Oran Gafni and Lior Wolf. 2020. Wish You Were Here: Context-Aware Human Generation. In *CVPR*. 7840–7849.
[11] W Keith Hastings. 1970. Monte Carlo sampling methods using Markov chains and their applications. (1970).
[12] Xin Jin, Le Wu, Xiaodong Li, Siyu Chen, Siwei Peng, Jingying Chi, Shiming Ge, Chenggen Song, and Geng Zhao. 2018. Predicting aesthetic score distribution through cumulative jensen-shannon divergence. In *AAAI Conference on Artificial Intelligence*.
[13] Xin Jin, Le Wu, Geng Zhao, Xiaodong Li, Xiaokun Zhang, Shiming Ge, Dongqing Zou, Bin Zhou, and Xinghui Zhou. 2019. Aesthetic Attributes Assessment of Images. In *Proceedings of the 27th ACM International Conference on Multimedia*. 311–319.
[14] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*. Springer, 662–679.
[15] Bert Krages. 2012. *Photography: the art of composition*. Simon and Schuster.
[16] Yining Lang, Wei Liang, Yujia Wang, and Lap-Fai Yu. 2019. 3d face synthesis driven by personality impression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1707–1714.
[17] Yining Lang, Wei Liang, and Lap-Fai Yu. 2019. Virtual agent positioning driven by scene semantics in mixed reality. In *IEEE VR*. 767–775.
[18] Yining Li, Chen Huang, and Chen Change Loy. 2019. Dense intrinsic appearance flow for human pose transfer. In *CVPR*. 3693–3702.
[19] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. 2019. Planercnn: 3d plane detection and reconstruction from a single image. In *CVPR*. 4450–4459.
[20] Ligang Liu, Renjie Chen, Lior Wolf, and Daniel Cohen-Or. 2010. Optimizing photo composition. In *Computer Graphics Forum*, Vol. 29. Wiley Online Library, 469–478.
[21] Wei Luo, Xiaogang Wang, and Xiaoou Tang. 2011. Content-based photo quality assessment. In *ICCV*. IEEE, 2206–2213.
[22] Shuang Ma, Yangyu Fan, and Chang Wen Chen. 2014. Pose maker: A pose recommendation system for person in the landscape photographing. In *Proceedings of the 22nd ACM international conference on Multimedia*. 1053–1056.
[23] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. 2018. Dense pose transfer. In *ECCV*. 123–138.
[24] Masashi Nishiyama, Takahiro Okabe, Imari Sato, and Yoichi Sato. 2011. Aesthetic quality classification of photographs based on color harmony. In *CVPR*. IEEE, 33–40.
[25] Ravi Ramamoorthi and Pat Hanrahan. 2001. An efficient representation for irradiance environment maps. In *SIGGRAPH*. ACM, 497–500.
[26] Norbert Schneider. 1994. *The art of the portrait: masterpieces of European portrait-painting, 1420-1670*. Taschen.
[27] Yujia Wang, Wei Liang, Jianbing Shen, Yunde Jia, and Lap-Fai Yu. 2019. A deep Coarse-to-Fine network for head pose estimation from synthetic data. *Pattern Recognition* 94 (2019), 196–206.
[28] Yujia Wang, Wenguan Wang, Wei Liang, and Lap-Fai Yu. 2019. Comic-guided speech synthesis. *TOG* 38, 6 (2019), 1–14.
[29] Wenyuan Yin, Tao Mei, Chang Wen Chen, and Shipeng Li. 2013. Socialized mobile photography: Learning to photograph with social context via mobile devices. *IEEE Transactions on Multimedia* 16, 1 (2013), 184–200.
[30] Lap-Fai Yu, Sai Kit Yeung, Demetri Terzopoulos, and Tony F. Chan. 2012. DressUp!: Outfit Synthesis Through Automatic Optimization. *TOG* 31, 6 (2012), 134:1–134:14.
[31] Yanhao Zhang, Xiaoshuai Sun, Hongxun Yao, Lei Qin, and Qingming Huang. 2012. Aesthetic composition represetation for portrait photographing recommendation. In *Proceedings of the 19th IEEE International Conference on Image Processing*. IEEE, 2753–2756.