Group Photo Realisation: A Survey of Multi person Group Photo Image Generation

Hongming Lu

Abstract— **Group photography is a common image in life, but when you take a group photo, there will always be someone who can't be there for a variety of reasons, resulting in a group photo always with regret. Especially in recent years, because of the epidemic, many people at the gathering could not be reached. As a result, the research on the generation of group photo images of people becomes more and more important. The purpose of this study is to perfectly embed a single person image into a multi-person photo image to make it harmonious and realistic. Previous people's image compositions focused more on the fusion of characters and backgrounds, with little regard to the interaction between people and other people in the group photo. Now, it is more desirable to have a good fusion with multi-person photo images. In this survey, we summarize the methods of the above research directions, and discuss the difficulties and future directions of future group photo character image generation to promote future research of group photo realization.**

Index Terms—**Group Photo, Image Generation, Posture Migration。**

I. INTRODUCTION

Since ancient times, human beings have been pursuing beauty, which has never changed. Therefore, in the field of image synthesis, the realization and harmony of images have been the pursuit of researchers. The purpose of image synthesis is to paste the foreground image from one image into another, combine different image elements, and build an image you want. We often use such operations in our daily life, generally through some image processing software to achieve their own purposes. However, there is nothing we can do to make the synthetic image more realistic and harmonious. For example, the popular "no way to see the world at home" is to synthesize single-person images in different countries and scenes.

The development of image synthesis[6,7] has further led to the study of group photo synthesis, which aims to synthesize a single foreground image into a group with multiple characters in order to achieve a true and harmonious effect. This brings some new challenges to image synthesis in a new field. For example, group photo character synthesis needs to consider the interaction of character postures and the harmony of character expressions. At the same time, some challenges still exist in the field of image synthesis, such as the consistency of light after people synthesis and the harmony of background fusion.

Manuscript received October 07, 2022.
Hongming Lu, School of Software, Tiangong University, Tianjin, China

II. RELATED WORK

The earliest research on group photo harmony [1], they proposed a framework to automatically enhance group photos through facial expression analysis. A character in the group photo may close his eyes inadvertently, may be looking elsewhere, or may not smile at this time. Their algorithm uses facial analysis to determine the goodness score of each facial instance in these photos. This scoring function trains a large number of annotated photos based on facial expression classifiers, such as smiling faces and closed eyes. Given these scores, you can (a) select a photo with the best overall score, and (b) use alignment and seamless mode to replace any low score faces in the photo with the high score faces of the same person from other photos to synthesize the best combination composition.

In addition, the study of group photography [2] proposed a new method to obtain visual balanced layout and dynamic visual elements. They used the concept of spring electric graphic model and expanded it with the concept of color energy in visual art literature. They also presented an interesting application of the proposed model in photographic aids. They focus on group photography and make recommendations to users using social media images and proposed spring powered models. The proposed method can provide users with real-time feedback on the arrangement of people, their position and the relative size on the image frame.

Until 2020, the research on group photos will gradually enter the right track. [3] This article proposes a new method to automatically synthesize virtual characters' postures according to users' postures, so that virtual characters can match each other from the perspective of visual aesthetics to form realistic and vivid photos. In order to take such pictures, the posture should be composed of appropriate angles of each body joint at the same time to meet the visual aesthetic standards. Compared with single person photography, in addition to the aesthetic expression of single person posture, double person photography also considers the connection and interaction between two people, especially the emotion. However, this study only considers the relationship between two people taking photos, and it is based on the research of virtual characters, which still has a certain distance from the real scene.

The real opening work in the research field of group photo synthesis was [4] of the Facebook research team, who proposed a new method to insert objects, especially people, into existing images, so that they can be mixed in a realistic way, while respecting the semantic context of the scene. Their method includes three subnets: the first is to generate a semantic map of a new person, the posture of other people in a

given scene, and optional bounding box specifications. The second network renders the pixels of novel characters and their hybrid masks based on the specification of multiple appearance components. The third network refines the generated face to match the face of the target person. Their experiment shows a convincing high resolution output in this novel and challenging application field. Although there are still some areas that can be improved in the aspect of character realism, this research does not affect its influence in the slightest.

Until now, on the basis of Facebook research, only one article [5] has improved its defects. They propose a novel method to generate context sensitive character images and insert them into existing scenes, while preserving global semantics. More specifically, our goal is to insert a person so that the position, posture and proportion of the person entering the scene can be integrated with the existing people in the scene. Their method uses three separate networks. First, predict the potential position and bone structure of new people through WGAN for existing human bones in the scene. Next, the predicted skeleton is refined through the shadow linear network to achieve higher structural accuracy in the generated image. Finally, another generation network is used to generate the target image from the refined skeleton. The network is based on the given image of the target person. In their experiments, they achieved photo realistic generation results with high resolution, while preserving the general background of the scene.

In general, the field of group photo synthesis is gradually developing, from the initial adjustment of the disharmonious expression in the group photo image to the consideration of interaction with other characters in the group photo. Maybe in the near future, the lighting factors in the group photo, the attribute factors of the characters themselves, and the inherent objects in the scene will be fully considered in the image synthesis until a perfect and harmonious group photo composite image is generated.

III. RESEARCH METHOD

In recent years, the research in the field of group photo image synthesis has been constantly exploring and developing. At present, the main research ideas and methods are divided into three categories, as detailed below:

A. Multi person posture aesthetics

The purpose of this method is to constrain the change of composite characters' posture from the perspective of aesthetic evaluation. In order to make the model have the aesthetic evaluation criteria of multi person pose, the pose skeleton corresponding to each person is generally extracted from the multi person group photo dataset through the pose estimation algorithm, and then the aesthetic evaluation network is used to learn the aesthetic expression of the multi person pose in the group photo..

Then, the trained aesthetic evaluation network of multi person pose is used to guide the pose of different characters in the multi person group photo image, so that they can be continuously optimized until an optimal pose is achieved. Next, the characters are rendered through the best pose of the group photo, so that the composite group photo image has a good pose harmony.

B. Prediction of multi person semantic interaction

The main purpose of this method is to predict the semantic analysis diagram of the inserted characters through the semantic analysis diagram of the fixed characters in the group photo, so that the position and posture of the newly inserted characters have good interaction and harmony with other characters in the group photo. The model training of this method is generally based on multi person analytic dataset MHP. A single semantic is randomly removed from the group photo, and then other multi person semantics are used as the input, and the real image semantics are used as the judgment. Through continuous network learning, the model has the ability to predict the semantics and location of new characters from group photos.

Next, after getting the semantic map and location of the new person, you need to render the semantic map for real people, but the rendered image may look unreal. Therefore, in the future, some detailed processing will be carried out on the face to make the image look clear and distinguishable.

C. Multi person attitude interaction prediction

The general idea of this method is basically consistent with the above semantic prediction, except that the purpose is to predict the posture skeleton and position of the newly inserted characters through the posture skeleton of the characters fixed in the group photo. In this way, we can better learn the interactivity of multi person posture through clear skeleton nodes. The training method is also basically the same as before, both of which are input by randomly removing the multi person pose skeleton after the single person, and the real multi person pose skeleton is used as the judgment.

Through the pre trained posture prediction model, the new person's posture is obtained, and then the real person's posture is transferred. Finally, the migrated real person is reasonably combined into the group photo image according to the predicted skeleton position.

Method	Idea	Advantages	disadvantages
A	The pose of the newly inserted characters in the group photo is predicted through the pretrained multi person pose aesthetic evaluation network.	The newly inserted character posture is more consistent with the aesthetic expression.	The method is not integrated, and each stage is relatively independent.
B	Predict the analytic diagram and position of newly inserted characters through the figure analytic diagram that is fixed in the group photo.	The newly inserted characters have a fixed position boundary, and have a good interaction with other group characters.	The subsequent rendering of characters on the semantic map will lead to unreal characters.
C	The pose skeleton and position of the newly inserted character are predicted by the fixed pose skeleton of the character in the group photo.	The position and posture of the newly inserted characters are more interactive with other characters in the group photo, and the posture is more clear.	Later, when the migrated character images are synthesized, there will be background disharmony.

Table 1

IV. COMPARISON OF METHODS

The above three methods are all exploration and research from different perspectives. Each method has its own emphasis, but also has its corresponding weaknesses. The above Table 1 clearly shows the advantages and disadvantages of these methods.

Although there are various methods, the data sets used for training these models are basically the same. MHP[10] datasets are generally used for multi person group photo image data sets, and DeepFashion[8,9] datasets are generally used for attitude data sets. This is also convenient for follow-up research to do comparative experiments.

V. CONCLUSION

In this paper, we conducted a comprehensive survey and summary of the field of multi person group photo image synthesis. This research belongs to a sub field of the image synthesis field. Although it is roughly the same as the image synthesis field, more problems and related technologies need to be considered, mainly involving the interaction of characters, the harmony of multiple characters, etc. In addition, we also summarized and discussed the advantages and disadvantages of the current research ideas and methods in this field, so as to facilitate future exploration and research based on these methods. At the same time, we hope that in the future, we can further consider the influence of people's lighting and the inserted new people's own attributes (height, body type, etc.) on the composite image in the field of group photo image synthesis.

REFERENCES

[1] Shah R, Kwatra V. All smiles: automatic photo enhancement by facial expression analysis. InProceedings of the 9th European conference on visual media production 2012 Dec 5 (pp. 1-10).
[2] Rawat YS, Song M, Kankanhalli MS. A spring-electric graph model for socialized group photography. IEEE Transactions on Multimedia. 2017 Sep 8;20(3):754-66.
[3] Wang Y, Hou S, Ning B, Liang W. Photo stand-out: Photography with virtual character. InProceedings of the 28th ACM International Conference on Multimedia 2020 Oct 12 (pp. 781-788).
[4] Gafni O, Wolf L. Wish you were here: Context-aware human generation. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020 (pp. 7840-7849).
[5] Roy P, Ghosh S, Bhattacharya S, Pal U, Blumenstein M. Scene Aware Person Image Generation through Global Contextual Conditioning. arXiv preprint arXiv:2206.02717. 2022 Jun 6.
[6] Weng S, Li W, Li D, Jin H, Shi B. Misc: Multi-condition injection and spatially-adaptive compositing for conditional person image synthesis. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020 (pp. 7741-7749).
[7] Zanfir M, Oneata E, Popa AI, Zanfir A, Sminchisescu C. Human synthesis and scene compositing. InProceedings of the AAAI Conference on Artificial Intelligence 2020 Apr 3 (Vol. 34, No. 07, pp. 12749-12756).
[8] Liu Z, Luo P, Qiu S, Wang X, Tang X. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 1096-1104).
[9] Ge Y, Zhang R, Wang X, Tang X, Luo P. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 5337-5345).
[10] Zhao J, Li J, Cheng Y, Sim T, Yan S, Feng J. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. InProceedings of the 26th ACM international conference on Multimedia 2018 Oct 15 (pp. 792-800).