Automatic Beautification for Group-photo Facial Expressions using Novel Bayesian GANs

Ji Liu[1], Shuai Li[1,2⋆], Wenfeng Song[1], Liang Liu[1], Hong Qin[3], and Aimin Hao[1]

[1]Beihang University, [2]Beihang University Qingdao Research Institute, [3]Stony Brook University (SUNY at Stony Brook)
{lishuai,ham}@buaa.edu.cn; qin@cs.stonybrook.edu

Abstract. Directly benefiting from the powerful generative adversarial networks (GANs) in recent years, various new image processing tasks pertinent to image generation and synthesis have gained more popularity with the growing success. One such application is individual portrait photo beautification based on facial expression detection and editing. Yet, automatically beautifying group photos without tedious and fragile human interventions still remains challenging. The difficulties inevitably arise from diverse facial expression evaluation, harmonious expression generation, and context-sensitive synthesis from single/multiple photos. To ameliorate, we devise a two-stage deep network for automatic group-photo evaluation and beautification by seamless integration of multi-label CNN with Bayesian network enhanced GANs. First, our multi-label CNN is designed to evaluate the quality of facial expressions. Second, our novel Bayesian GANs framework is proposed to automatically generate photo-realistic beautiful expressions. Third, to further enhance naturalness of beautified group photos, we embed Poisson fusion in the final layer of the GANs in order to synthesize all the beautified individual expressions. We conducted extensive experiments on various kinds of single-/multi-frame group photos to validate our novel network design. All the experiments confirm that, our novel method can uniformly accommodate diverse expression evaluation and generation/synthesis of group photos, and outperform the state-of-the-art methods in terms of effectiveness, versatility, and robustness.

Keywords: Beautification of Group-photo Facial Expressions; Multi-label CNN; Bayesian Networks; Generative Adversarial Networks; Poisson Fusion.

1 Introduction and Motivation

With the omnipresence of digital cameras in today's society, group photos are routinely captured to record wonderful moments shared by families, friends, colleagues, etc. Hence, higher expectations are focused on the overall quality of group photos. In practice, it is almost impossible to capture satisfying facial

⋆ Corresponding author: Shuai Li

expressions in a synchronous way for all involved people at any moment with various types of hand-held devices. Therefore, it urgently needs to develop smart group photo evaluation and beautification techniques. However, to achieve this goal, there are still several challenges yet to be overcome, including evaluation of the group-photo facial expression on an individual basis, simultaneous generation of satisfying expressions for all people involved, natural synthesis integration of individual facial expression into the final production of a group photo, etc. Obviously, evaluation and beautification of facial expressions in such unconstrained settings remain an ill-posed task due to various factors, such as non-frontal faces, varying lighting in different outdoor/indoor settings, and/or even the large variation in facial identities and appearances.

With a goal of tackling the aforementioned challenges, more research works began to endeavor great efforts in related techniques. For example, recent works have demonstrated generative adversarial networks (GANs) are extremely effective. This ranges from image translation [8, 20, 17, 6], to face generation [16, 15, 13, 2] and even image completion [7, 4, 12]. Nonetheless, most of the existing methods commonly employ the entire feature space to approximate the generative feature distribution, which could not well respect facial expression details for all individuals involved. In addition, most of the existing works concentrate on the attribute manipulation/transformation of single object, lacking a principled way to optimize group-photo facial expressions.

In this paper, our research efforts are devoted to pioneering a systematic approach for synthesizing a satisfying group photo by leveraging the synchronized power of CNNs and GANs. Specifically, we propose a two-stage deep network for automatic group-photo evaluation and beautification, which could greatly reduce the negative influences caused by the diversity of faces. Fig. 1 highlights the framework of our novel method, which mainly consists of three major steps: (1) Facial expression recognition with multi-label CNN and our newly-proposed facial expression evaluation metric — the multi-label CNN recognizes two main beautification related expressions (e.g., mouth-smiling and eyes-opening) and predicts the softmax value of the expression for further evaluation; (2) Face beautification with our Bayesian GANs — it is guided by the subspace clustering based on attributes-aware priors, wherein we pre-distribute all the attributes' weights according to the specific face regions' impacts on the entire face appearances; (3) Multiple single-person faces' integration driven by ensemble Poisson fusion — we add a Poisson layer to naturally fuse single-person face into the original group photo with gradual gradient changes. The salient contributions of this paper can be summarized as follows:

- We pioneer a two-stage group-photo beautification framework by combining multi-label CNN with Bayesian network enhanced GANs, which could naturally and automatically perform evaluation and beautification on group photos in a uniform and elegant way.
- We propose novel Bayesian GANs to automatically generate beautiful expressions by embedding Bayesian prior network into the powerful Cycle-

GANs, which has strong generalization ability for weakly-matched training datasets.

– We propose to embed the Poisson image clone technique in the final layer of our Bayesian GANs in order to synthesize all the to-be-beautified expressions on all individuals from single-/multi-frame continuous group photos, which would lead to meaningful and harmonious manipulation in any local region of a group photo.

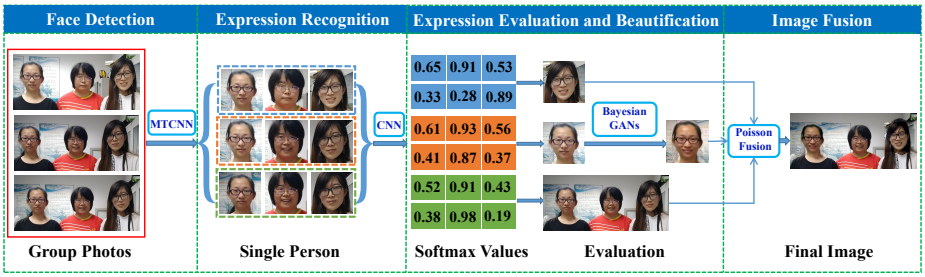

Fig. 1. The architecture of our framework. A group photo is converted into several single-person faces by using the MTCNN [19], which is a multi-task cascaded convolutional network to process the face detection tasks from coarse to fine.

2 Related Works

Facial Expression Recognition Methods. Facial expression recognition has been gaining growing momentum, with a wide range of applications. Specially, the expression recognition methods based on CNNs [1, 18] and DBN [9] have achieved excellent results on facial datasets. For example, Burkert et al. [1] proposed a facial emotion recognition architecture based on CNNs. It consists of two parallel feature extraction blocks (FeatEx), which dramatically improves the performance on public datasets. Liu et al. [9] proposed a boosted deep belief network (BDBN) for feature learning, feature selection, and classification in a loopy framework. However, these methods are in some sense cumbersome due to high-dimensional varying features for each attribute, leading to inefficiency in recognition. Therefore, we apply multi-task learning to simultaneously optimize multiple objective functions.

Facial Expression Generation and Editing Methods. In recent years, many image generation approaches have been proposed. For example, Isola et al. proposed a pix2pix approach [5] and achieved amazing results on paired datasets. However, in many cases, paired data are not readily available. Therefore, the image conversion based on unpaired data is particularly important. Recently, Zhu et al. proposed the CycleGAN [20] method, which employed two GANs and an additional cycle consistency loss to improve the quality of the generated images. Meanwhile, DualGAN and DiscoGAN [17, 6] adopted the similar idea for image-to-image translation based on unpaired data. Particularly, many GAN-based methods have also been proposed for face generation. Perarnau et

al. introduced ICGAN [13], which combined the encoder with cGAN to manipulate face images conditioned on arbitrary attributes. Shen et al. introduced a framework [15] to avoid learning redundant facial information by learning residual images, which only focused on the attribute-specific area of a face image. However, these works commonly have significant dependencies on the training dataset and are difficult to preserve more details on other images. Moreover, these methods are designed for single pre-processed face images instead of group photos. Therefore, we should solve this to achieve strong generalization ability for weakly-matched test datasets.

3 Facial Expression Evaluation and Beautification

3.1 Facial Expression Evaluation based on Multi-label CNN

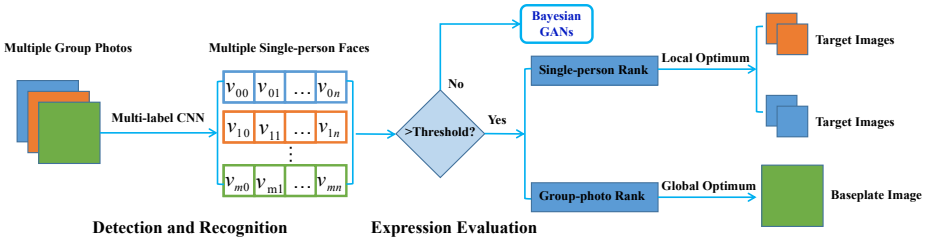

Fig. 2. Illustration of our facial expression evaluation pipeline.

In order to synthesize group photo with perfect facial expressions, we need to first select the face images that will be manipulated after face detection. Considering the unbalanced distribution of samples in the training and testing phases for multi-label classification, we adopt a mixed objective optimization network [14] to recognize different facial attributes. We perform a joint optimization over all the face attributes on CelebA dataset [10]. In practice, we focus on two main beautification related attributes, including mouth-smiling and eyes-opening. Based on the two attributes, we further construct a multi-label CNN to recognize the two expressions at the same time, and this multi-task loss is defined as

$$L(x, y) = \sum_{i=1}^{2} p(i|y_i(x)) \|f_i(x) - y_i(x)\|^2, \tag{1}$$

where $p(i|y_i(x))$ is the assigned probability for the attribute i, which can make the training set biased. $f_i(x)$ and $y_i(x)$ respectively represent the predicted value and the ground truth for attribute i. Meanwhile, we formulate a beautification evaluation metric for facial expressions, which facilitates beautifying group photos with lower cost. First, we count the number of individual faces with better expression in each group photo, so that we can choose the relatively better group photo to serve as our baseplate image. The metric used for measuring facial expression is the softmax value $V_i = e^{z_i} / \sum_{j=1}^{2} e^{z_j}$, which is obtained from the recognition network. As shown in Fig. 2, the softmax value v_{mn} means the n-th

person of the m-th group photo. We can directly substitute the target image with the highest softmax value (the softmax value must be greater than 0.5) for the worse one (the softmax value is less than 0.5) in the baseplate image using our improved Poisson fusion. It should be noted that, if there is no satisfying facial image of certain person, we resort to our Bayesian GANs to generate a desirable image.

3.2 Facial Expression Beautification with Bayesian GANs

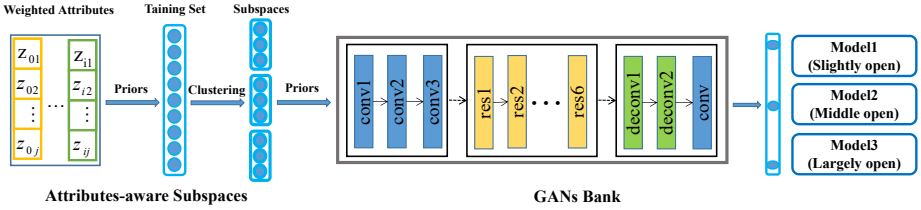

Fig. 3. Pipeline of our Bayesian GANs based on facial-attribute priors.

Considering the importance of diverse faces with various kinds of attributes, as shown in Fig. 3, we propose a three-layer Bayesian network to augment GAN models. Of which, the first layer of the Bayesian network relates to the attributes distribution prior, which is vital to cluster the semantics-similar images into one attributes-specific subspace. The second layer relates to the subspaces, which are clustered according to the attributes' influences on the targeted face regions. The third layer relates to the trained GANs, which are guided by the attributes-aware priors resulted from subspace clustering.

In the first layer, we pre-distribute the attributes' weights according to the specific regions' impacts on the entire face appearance. The j-th original attribute label value of the i-th sample $z_{ij} \in \{1, -1\}$ is re-distributed to $z_{ij} \in \{a_{ij}, 0\}$. Of which, a_{ij} denotes the new weight of the positive attribute value, and the '0' means the negative attribute value, which has no effects on face appearances. Based on such re-weighted attribute distribution in the first layer, we employ the k-means algorithm to perform subspace clustering on the training images according to the diverse attributes' influence on the targeted face regions. Here, we use the mean square errors of the attribute vectors to cluster all the samples into K subspaces,

$$E = \sum_{i=1}^{K} \sum_{z \in \mathcal{S}_i} \|z - u_i\|^2, \tag{2}$$

where $\mathcal{S}_i$ denotes the i-th subspace, and $\|z - u_i\|^2$ is the Euclidean distance between sample z and the subspace center u_i.

After attribute-aware subspace clustering, we further describe the image sample generating process from source domain X to target domain Y in details. Given two datasets X, Y: source domain $X = \{x_i | 1 \leqslant i \leqslant n_x\}$ and target domain $Y = \{y_i | 1 \leqslant i \leqslant n_y\}$, n_x, n_y respectively represent the numbers of dataset X

and Y. We cluster the sample space into three subspaces $\mathcal{S}_i, i = 1, 2, 3$ based on the attributes with important impacts. With the mapping function $X \to Y$, we adopt a loss function as:

$$L_{X\to Y}^{\mathcal{S}_i} = \mathbb{E}_{y\sim p_{data}(y)}[(D_Y(y) - 1)^2] + \mathbb{E}_{x\sim p_{data}(x)}[(1 - D_Y(G(x)))^2], \quad (3)$$

where $X, Y \in \mathcal{S}_i$. Therefore, our Bayesian GANs have excellent generation ability, which can successfully transform images between two domains according to the attribute-specific subspaces. Considering a test image, we first predict its 40 facial attributes using a multi-label CNN model, and then calculate which subspace the test image belongs to, according to the prior knowledge and the Bayesian network.

For our generator, we use three convolution layers to extract features from input images, six residual blocks to preserve the features of the original image, and simultaneously transform feature vectors from source domain to target domain. Meanwhile, we use three deconvolution layers to restore low-level features from feature vectors. Residual blocks consist of two convolution layers, wherein part of the input data is directly added to the output, so that we can reduce the deviation of the corresponding output from the original input. Finally, we use four convolution layers for our discriminative network.

3.3 Poisson Fusion in our GANs

To obtain a natural group photo, we need to conduct global fusion via local image editing [11]. Therefore, we embed a Poisson fusion layer in our GANs' final layer. In this layer, we naturally fuse all the generated facial expressions of different persons into the selected baseplate group photo. The key of Poisson fusion is to obtain the transformed pixel by solving the Poisson equation. Here, we construct the linear equation according to the method of Poisson image editing as: $\boldsymbol{A} \times \boldsymbol{x} = \boldsymbol{b}$. Please refer to [3] for the details about this equation.

If we solve the above Poisson equation with Gaussian elimination, it will exhibit a lot of time and memory cost. Considering the fusion region is a rectangle, some characteristics of matrix $\boldsymbol{A}$ can be leveraged: $\boldsymbol{A}$ is sparse, positive definite, and can be partitioned into smaller square matrices. According to these characteristics, we adopt the conjugate gradient method to solve the equation. And we do not need to store the matrix $\boldsymbol{A}$, because the conjugate gradient method only needs the value of $\boldsymbol{A} \times \boldsymbol{p}$, which can be easily obtained via the operation of block matrix. Thus, our method not only can embed larger region, but also can achieve more than 5000 times speedup (compared to the Gaussian elimination method) when both the height and width of the region are 100 pixels.

In practice, for ease of image synthesis, we need to store the facial coordinate information during face detection. By means of Poisson fusion method, the generated target images can be seamlessly fused into the selected baseplate group photo. Meanwhile, it can well keep the consistency of the color, texture, and illumination in the scene.

4 Experimental Results and Evaluations

Experimental Settings. We carefully design three types of experiments to evaluate the overall performance of our method: (1) single-person facial expression beautification of a group photo; (2) single-frame image based group photo beautification (the images are randomly-crawled from the internet); (3) multi-frame continuous images based group photo beautification (the images are captured by our hand-held device). CelebA is used as our training dataset, which includes 202,599 colored face images and 40-attribute binary vectors for each image. We use the aligned and cropped version and scale the images to the size of 128×128. In addition, the distribution of attribute labels are highly biased. In practice, for each attribute that needs to be edited, 1000 images from the attribute-positive class and 1000 images from the attribute-negative class are randomly chosen as our test set. We select all the rest images as our training dataset. Meanwhile, to demonstrate the superiorities of our method, we randomly search some facial images from the internet and take some photos casually, which also serve as our test dataset. Please refer to our supplemental document for more vivid results[1].

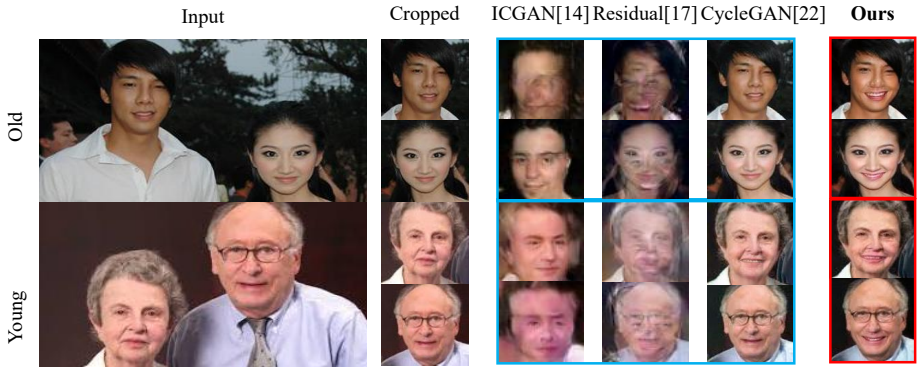

Fig. 4. Comparison of the mouth-smiling results produced by different methods on single-person faces of a group photo.

Evaluations on Single-person Facial Expression Beautification of Group Photos. Considering the detailed wrinkles on elder faces, we respectively conduct experiments on the different-age faces of group photos. As shown in Fig. 4, we compare our results with those produced by some state-of-the-art methods, including ICGAN [13], learning residual images [15], and CycleGAN [20]. We observe that, the compared methods commonly have a significant dependence on the training dataset, thus, their results on other test images are not satisfactory. In sharp contrast, our results are more natural and can preserve more details. Moreover, when facing diversified and complicated expression manipulation tasks, our approach outperforms the state-of-the-art facial expression beautifying methods with respects to effectiveness, versatility, and robustness.

Evaluations on Single-frame Image based Group Photo Beautification. In this kind of experiments, we use our generalization network to manipu-

[1] https://drive.google.com/file/d/159my8s52wzL-Eq9vGtubKDegMQLfLfQq/view?usp=sharing

late facial attributes and further synthesize a beautiful group photo. Our network can successfully synthesize semantically-meaningful and visually-plausible contents for the key face regions that need to be beautified. As shown in Fig. 5, our method can generate satisfying results with high perceptual quality, which shows a great promise for smart facial expression beautification during group photo capturing.

Fig. 5. The results of our method for single-frame image based group photo beatification.

Evaluations on Multi-frame Continuous Images based Group Photo Beautification. Our method can also synthesize a new satisfying group photo from unsatisfying multi-frame continuous images. Considering diverse poses in multi-frame continuous group photos, we detect facial landmarks from the generated images and a group photo to locate a rectangle region of eyes/mouth for ease of fusing the manipulated regions. As shown in Fig. 6, we replace worse facial expressions with the beautified ones in the baseplate group photo based on our improved Poisson fusion strategy.

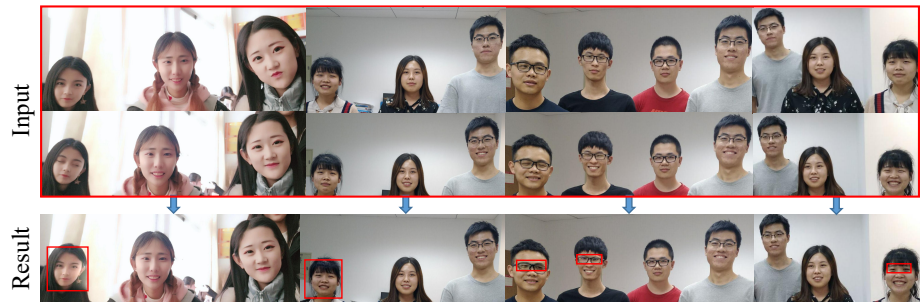

Fig. 6. The results of our method for multi-frame continuous images based group photo beautification.

Quantitative Evaluations. To quantitatively evaluate the visual quality of the synthesized group photos, we carry out user study, wherein 20 people are asked to classify the randomly shuffled images as real or synthetic ones. Each person is shown a random selection of 50 real images and 50 synthesized images in a random order, and is asked to label the images as either real image or synthetic image. Table 1 shows the confusion matrix, which indicates that, people feel very hard to reliably distinguish real images from our synthetic ones.

Meanwhile, we conduct user study based on the survey from 20 participants, wherein participants are required to assess the visual realism, image quality, and individual detail preservation by asking them to label the best generated image from the randomly shuffled images generated by different methods. Table 2

documents the results. For the voting about the best performance on attributes manipulation, our method gains the majority of votes. It clearly shows that, our method can well accommodate photo-realistic facial expression beautification for highly-diverse group photos. In addition, as shown in Fig. 7, we further ask participants to grade our results between 0 and 5 according to the image quality and visual realism. It confirms that, our method outperforms other approaches on facial expression beautification.

	Labeled As Real	Labeled As Synthetic
Real	557	443
Synthetic	412	588

Table 1. Visual Turing test results for distinguishing real/synthesized images. The average human classification accuracy is 57.25% (chance = 50%).

Methods	Mouth-Smiling Beautification	Eyes-Opening Beautification
ICGAN	0.7%	-
Residual	2.1%	1.3%
CycleGAN	37.3%	45.4%
Ours	**59.9%**	**53.3%**

Table 2. Visual Turing test results about different-methods' facial expression manipulations on group photos. The voting percentage sum of each column is equal to 100%.

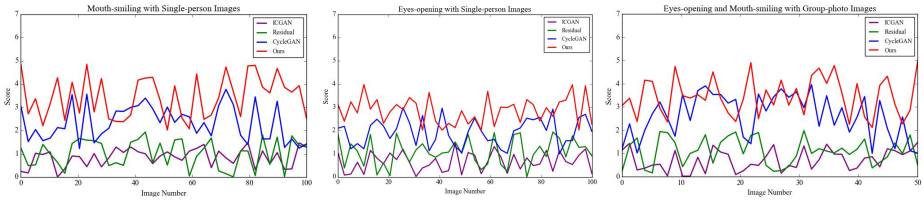

Fig. 7. The subjective evaluation on different methods.

5 Conclusion and Future Works

This paper detailed a two-stage first-evaluation-then-beautification framework with which we could synthesize satisfactory group photos from original single- or multi-frame group photos that are routinely-captured in our daily life. Benefiting from the novel integration of multi-label CNN and Bayesian prior embedded GANs, our novel framework could generate natural and realistic images, which helps improve the generalization ability of facial expression manipulation and synthesis. Various qualitative and quantitative experiments were carried out to evaluate the overall performance of our method, and all the experiments confirmed that, our method has apparent advantages over the existing techniques in terms of efficacy, effectiveness, versatility, and robustness. Despite many promising results in most cases, the obtained results are sometimes less ideal. For example, it remains difficult to generate and synthesize group photos when we only have images of low quality, or images involving facial occlusion and/or complex body pose. Such challenging cases deserve more research efforts. Besides, we plan to exploit more intrinsic temporal context priors and how such priors could further enhance group photo beautification in the near future.

6 Acknowledgments

This research is supported in part by National Natural Science Foundation of China (NO. 61672077 and 61532002), Applied Basic Research Program of Qingdao (NO. 161013xx), National Science Foundation of USA (NO. IIS-0949467, IIS-1047715, IIS-1715985, and IIS-1049448), National Key R&D Program of China (NO. 2017YFF0106407), and capital health research and development of special 2016-1-4011.

References

1. Burkert, P.e.a.: Dexpression: Deep convolutional neural network for expression recognition. arXiv preprint arXiv:1509.05371 (2015)
2. Cole, F., Belanger, D., Krishnan, D., Sarna, A., Mosseri, I., Freeman, W.T.: Synthesizing normalized faces from facial identity features. In: CVPR (2017)
3. Gangnet, M., Blake, A.: Poisson image editing. In: SIGGRAPH. pp. 313–318 (2003)
4. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and Locally Consistent Image Completion. TOG 36(4), 107:1–107:14 (2017)
5. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
6. Kim, T., Cha, M., et al., H.K.: Learning to discover cross-domain relations with generative adversarial networks. In: ICML. vol. 70, pp. 1857–1865 (2017)
7. Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. In: CVPR (2017)
8. Liu, M., Tuzel, O.: Coupled generative adversarial networks. NIPS pp. 469–477 (2016)
9. Liu, P., Han, S., Meng, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: CVPR (2014)
10. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
11. Mccann, J., Pollard, N.S.: Real-time gradient-domain painting. SIGGRAPH 27(3), 93 (2008)
12. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR (2016)
13. Perarnau, G., van de Weijer, J., Raducanu, Bogdan, j.y.: Invertible conditional gans for image editing
14. Rudd, E.M., Gunther, M., Boult, T.E.: Moon: A mixed objective optimization network for the recognition of facial attributes. ECCV pp. 19–35 (2016)
15. Shen, W., Liu, R.: Learning residual images for face attribute manipulation. In: CVPR (2017)
16. Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., Samaras, D.: Neural face editing with intrinsic image disentangling. In: CVPR (2017)
17. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: ICCV (2017)
18. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: ICMI. pp. 435–442 (2015)
19. Zhang, K.e.a.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters 23(10), 1499–1503 (2016)
20. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)