




# Multi Facet Face Construction

Hamed Alqahtani<sup>1</sup>(✉)  and Manolya Kavakli-Thorne<sup>2</sup>

<sup>1</sup> King Khalid University, Abha, Saudi Arabia

hsqahtani@kku.edu.au

<sup>2</sup> Macquarie University, Sydney, Australia

**Abstract.** To generate a multi-faceted view, from a single image has always been a challenging problem for decades. Recent developments in technology enable us to tackle this problem effectively. Previously, Several Generative Adversarial Network (GAN) based models have been used to deal with this problem as linear GAN, linear framework, a generator (generally encoder-decoder), followed by the discriminator. Such structures helped to some extent, but are not powerful enough to tackle this problem effectively.

In this paper, we propose a GAN based dual-architecture model called DUO-GAN. In the proposed model, we add a second pathway in addition to the linear framework of GAN with the aim of better learning of the embedding space. In this model, we propose two learning paths, which compete with each other in a parameter-sharing manner. Furthermore, the proposed two-pathway framework primarily trains multiple sub-models, which combine to give realistic results. The experimental results of DUO-GAN outperform state of the art models in the field.

**Keywords:** GAN · Multi-faceted face construction · Neural network · Machine learning

## 1 Introduction

Constructing a multi-faceted image from a single image is a well-investigated problem and has several real-life applications. Essential applications of creating a multi-posed image from a single image are its use for identification purposes, detecting malicious, criminals in public, capturing the identity of people in general etc. Constructing multi-posed image is a challenging task comprising of imagining the objects might looking like, constructed from another pose [3]. It requires the construction of unknown possibilities and hence requires a very rich embedding space so that the constructed view of the object should have the same identity and should be relevant in context.

Several research efforts have been made to address this problem using different models like synthesis based models, and data-based models [16, 19]. These GAN based models consist of linear framework and encoder-decoder followed by Discriminator to address this issue. Here, the main purpose of the encoder(E) is to map the input images to the latent space(Z), which are fed into the decoder(G) after some manipulation for generating multi-faceted images [1, 2].

But, it is found empirically that the linear framework isn't powerful enough to learn appropriate embedding space. The linear framework generates an output for creating a multi-faceted image isn't clear enough and doesn't preserve identity across various posed images. Learning incomplete embedding space leads to incomplete generalization on test images or unseen images. The primary reason of incapability of linear frameworks in learning complete presentation is that during training the encoder part of  $G$  only sees a fraction of  $Z$  and while testing, very likely model come across samples corresponding to unseen embedding space. This results in poor generalization.

In order to tackle this problem, Tian et al. [14] proposed a dual-pathway architecture, termed as Complete-Representation (CR-GAN). Unlike linear framework, the authors of CR-GAN have used dual pathway architecture. Besides the typical re-construction path, they introduced another generation path for constructing multi-faceted images from embeddings, randomly sampled from  $Z$ . In the proposed architecture, they used the same  $G$ , which aids the learning of  $E$  and discriminator ( $D$ ). In their proposed model,  $E$  is forced to be an inverse of  $G$ , which theoretically should yield complete representations that should span the entire  $Z$  space.

However, the experiments conducted in this work demonstrate that one encoder is not convincing to span the entire  $Z$  space. Therefore, in order to address this challenge, we propose DUO-GAN with *dual encoder* to learn complete representation for a multi-facet generation. The primary purpose is to distribute the task of spanning the entire  $Z$  space, across two encoders instead of one as proposed in the previous work. We empirically demonstrate that dual encoder architecture produces many realistic results in comparison to prior work in this field.

## 2 Related Work

Several researchers contributed to constructing a multi-faceted image from a single image. The significant work in this field is presented as follows.

Goodfellow et al. [5] first introduced GAN to learn models with generative ability via an adversarial process. In the proposed model, a two-player min-max game is played between generator ( $G$ ) and discriminator ( $D$ ). Competing with each other in the game, both  $G$  and  $D$  tend to improve themselves. GAN has been used in various fields like image synthesis, super-resolution image generation etc. Every model proposed with the help of GAN manipulates constraints on  $Z$  and attempt to cover more and more embedding space for a better synthesis of images.

Hassner et al. [8] proposed a 3D face model in order to generate a frontal face for any subject. Sagonas et al. [13] used a statistical model for creating joint frontal face reconstruction, which is quite useful. The reported results were not very useful, as frontal face generation from a side view is a very challenging task. Because of occlusion and variation in spatial feature from side view face pictures.

Yan et al. [16] solved the problem of multi-pose generation to a certain level by using projection information by their Perspective Transformer Nets. Whereas, Yang et al. [17] proposed a model which incrementally rotated faces in fixed yaw angles. For generating multi-poses, Hinto et al. [9] tried generating images with view variance by using auto-encoder. Tian et al. [14] proposed dual pathway architecture CR-GAN for constructing multiple poses. However, all the above-mentioned system fail to construct realistic images in an unseen wild condition. In comparison, DUO-GAN spans embedding space in a much more exhaustive manner using its multi-path architecture and produces higher-quality images than previously proposed models.

Preserving identity synchronously across images with numerous positions is a very active research area. Previously DR-GAN [15] attempted to solve this problem, by providing pose code along with image data, while training. Li et al. [12] attempted this challenge by using *Canonical Correlation Analysis* for comparing the difference between the sub-spaces of various poses. Tian et al. [14] tried solving this problem with dual pathway architecture. We propose dual encoder dual-pathway architecture, which results in a much better generation of multi-faceted images.

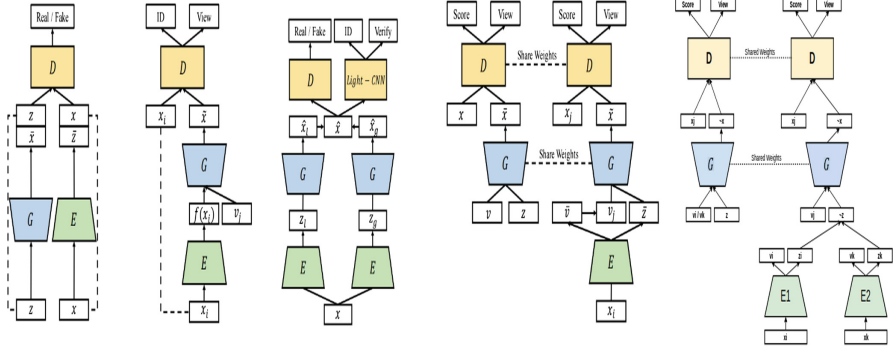
### 3 The Proposed Method

Most of the previous research on this field involves a linear network, *i.e.* an encoder-decoder generator network, followed by Discriminator network. As empirically found, such linear network is incapable of spanning entire embedding space, which leads to incomplete learning as a single encoder can only span limited space, irrespective of the variance and quantity of data. So while testing, when an unseen image is passed through the G, it is very likely that the unseen input will be mapped to un-covered embedding space, which consequently leads to the poor generation of images.

Yu et al. [14] proposed CR-GAN, which uses dual-pathway architecture to cover embedding space more extensively than a linear framework. It's primary uses a second-generation path, with the aim to map the entire  $Z$  space to corresponding targets. However, we empirically found that single encoder used in dual pathway architecture is not powerful enough to span the entire embedding dimension. This fact motivates us to use dual encoder architecture for spanning embedding space more extensively. Figure 1 illustrates the comparison between our proposed model, CR-GAN and other linear networks. The proposed model consists of two paths, namely Generator path, and Reconstruction path, described in following subsections.

#### 3.1 Generator Path

This path is similar to the Generator path proposed in CR-GAN [14]. Here both the encoder are not involved, and G is trained to generate with random noise. Here we give a view-label  $v$  and random noise  $z$ . Aim is to produce very realistic



**Fig. 1.** Comparison of models: BiGAN, DR-GAN, TP-GAN, CR-GAN, and the proposed model

image  $G(v, z)$  with view-label  $v$ . And like in GANs aim of  $D$  is to distinguish the output of  $G$ 's from real.  $G$  tries to minimize Eqs. 1 and 2.

$$\mathbb{E}_{\mathbf{z} \sim \mathbf{P}_{\mathbf{z}}} [D_s(G(v, \mathbf{z}))] - \mathbb{E}_{\mathbf{x} \sim \mathbf{P}_{\mathbf{x}}} [D_s(\mathbf{x})] + \mathbf{C}_1 \mathbb{E}_{\hat{\mathbf{x}} \sim \mathbf{P}_{\hat{\mathbf{x}}}} [(\|\nabla_{\hat{\mathbf{x}}} \mathbf{D}(\hat{\mathbf{x}})\|_2 - 1)^2] - \mathbf{C}_2 \mathbb{E}_{\mathbf{x} \sim \mathbf{P}_{\mathbf{x}}} [\mathbf{P}(\mathbf{D}_{\mathbf{v}}(x) = \mathbf{v})] \quad (1)$$

Here,  $\mathbf{P}_{\mathbf{x}}$  represents the distribution of data, and  $\mathbf{P}_{\mathbf{z}}$  represents the uniform noise distribution. Further,  $\mathbf{P}_{\hat{\mathbf{x}}}$  represents the interpolation between the data constructed from different images. In the proposed model, we randomly pass either  $v_i$  or  $v_k$ , as we want to learn  $G$  to generate high quality images either from  $\hat{\mathbf{x}}$  which is interpolation of  $x_i$  and  $x_k$  as further discussed in Sect. 3.2. We also experimentally found that feeding in  $\hat{\mathbf{x}}$  in first phase of training did not give good results, possibly because of noise, formed due to interpolation.

$$\mathbb{E}_{\mathbf{z} \sim \mathbf{P}_{\mathbf{z}}} [D_s(G(v, \mathbf{z}))] + \mathbf{C}_3 \mathbb{E}_{\mathbf{z} \sim \mathbf{P}_{\mathbf{z}}} [P(D_v(G(v, \mathbf{z})) = v)] \quad (2)$$

The proposed algorithm for training our model in phase 1 and phase 2, with batch-size  $b$  and time-steps  $t$  is described as below.

---

## Algorithm

---

**Input:** Sets of images  $X$ .

**Result:** Trained architecture,  $G$ ,  $D$ ,  $E1$ ,  $E2$ .

1. Sample  $\mathbf{z}_1 \sim \mathbf{P}_{\mathbf{z}}$ ,  $\mathbf{x}_i \sim \mathbf{P}_{\mathbf{x}}$  with  $\mathbf{v}_i$  and  $\mathbf{x}_k \sim \mathbf{P}_{\mathbf{x}}$  with  $\mathbf{v}_k$ ;
  2.  $\hat{\mathbf{x}} \leftarrow \mathbf{G}(\mathbf{v}_i, \mathbf{z}) \mathbf{n} \mathbf{G}(\mathbf{v}_k, \mathbf{z})$ ;
  3. Update  $D$  by Eq. 1, and  $G$  with Eq. 2;
  4. Sample  $\mathbf{x}_j$  with  $\mathbf{v}_j$  (where  $\mathbf{x}_j$ ,  $\mathbf{x}_i$  and  $\mathbf{x}_k$  share the same identity);
  5.  $(\hat{\mathbf{v}}_i, \hat{\mathbf{z}}_i) \leftarrow \mathbf{E}_1$ ;
  6.  $(\hat{\mathbf{v}}_k, \hat{\mathbf{z}}_k) \leftarrow \mathbf{E}_2$ ;
  7.  $\hat{\mathbf{x}}_j \leftarrow G(\mathbf{v}_j, \mathbf{z})$ ;
  8. Updated  $D$  by Eq. 3, and  $E$  by Eq. 4;
-

### 3.2 Reconstruction Path

We train both the **E1** and **E2** and **D** but not the **G**. In reconstruction path we make **G** generate image from the features extracted from **E1** and **E2** re-generate images, which makes them both inverse of **G**. Passing different poses in both **E1** and **E2** makes sure they cover different embedding space, which in turns leads to complete learning of latent embedding space. Further, the output generated from the **E1** and **E2** is combined using the interpolation between the data points from each of encoders, which are in spirit the same as  $\hat{x}$  in generation part.

For making sure the re-constructed images by **G** from the features extracted from **E1** and **E2** share the same identity we use the cross reconstruction task, in order to make **E1** and **E1** preserve identity. To be more precise, we pass in image of same identity in both **E1** and **E2** having different poses. As primary goal is to re-construct an image  $\mathbf{x}_j$  with interpolation of images  $\mathbf{x}_i$  and  $\mathbf{x}_k$ . So in order to do this, **E1** takes  $\mathbf{x}_i$  and **E2** takes  $\mathbf{x}_k$ , both of these encoders output an identity preserved  $\bar{\mathbf{z}}_i$  and  $\bar{\mathbf{z}}_k$  with respective view estimation  $\bar{\mathbf{v}}_i$  and  $\bar{\mathbf{v}}_k$ .

**G** takes  $\bar{\mathbf{z}}$  and view  $\mathbf{v}_i$  as input, and is trained to reconstruct the image of the same person with view  $\mathbf{v}_i$  with the help of interpolated  $\bar{\mathbf{z}}$ . Here  $\bar{\mathbf{z}}$  should help **G** to preserve identity and carry out essential latent features of the person. **D** here is trained to differentiate between the fake image  $\hat{\mathbf{x}}_j$  from the real one  $\hat{\mathbf{x}}_i$  or  $\hat{\mathbf{x}}_k$ . Thus, **D** minimizes the Eq. 3.

$$\begin{aligned} \mathbb{E}_{\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k \sim \mathbb{P}_{\mathbf{x}}} [2 \times D_s(\hat{x}_j) - D_s(x_i) - D_s(x_k)] + \mathbf{C}_1 \mathbb{E}_{\hat{\mathbf{x}} \sim \mathbb{P}_{\hat{\mathbf{x}}}} [(\|\nabla_{\hat{\mathbf{x}}} \mathbf{D}(\hat{x})\|_2 - 1)^2] \\ - \mathbf{C}_2 \mathbb{E}_{\mathbf{x}_i \sim \mathbb{P}_{\mathbf{x}}} [\mathbf{P}(\mathbf{D}_{\mathbf{v}}(x_i) = \mathbf{v}_i)] \end{aligned} \quad (3)$$

Here,  $\tilde{\mathbf{x}} = \mathbf{G}(\mathbf{v}_j, \mathbf{E}_z(\mathbf{x}_i))$ . **E** helps **G** to generate realistic image, with  $\mathbf{v}_j$ . Basically, **E1** and **E2** maximizes Eq. 4.

$$\mathbb{E}_{\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k \sim \mathbb{P}_{\mathbf{x}}} [D_s(\hat{x}_j) + \mathbf{C}_3 \mathbf{P}(\mathbf{D}_{\mathbf{v}}(\tilde{x}_j) = \mathbf{v}_j) - \mathbf{C}_4 \mathbf{L}_{j1}(\tilde{x}_j, \mathbf{x}_j) - \mathbf{C}_5 \mathbf{L}_v(\mathbf{E}_v(\mathbf{x}_i), \mathbf{v}_i)] \quad (4)$$

Here,  $\mathbf{L}_1$  is the loss to ensure  $\tilde{x}_j$  is reconstructed property from  $\mathbf{x}_j$ .  $\mathbf{L}_v$  is the loss estimated from cross-entropy of the ground and estimated views, for **E1** and **E2**.

This dual-dual-pathway network efficiently spans complete embedding space. In the first path of the algorithm, **G** learns how to better produce image, from the random noise, which in time, when produced through the **E1** leads to better output.

In comparison to previously proposed linear-networks, the proposed double-dual pathway network helps better solve the problem of multi-facet construction in following ways:

- It leads to better covering of latent embedding space, which in turns leads to better generation of multi-faceted pictures.
- Once trained on good quality images, model seems to work pretty well even for low quality images, probably because of expansive embedding space covered.

## 4 Experiments and Results

This section describes the experimental setup, benchmark dataset, experimental results and compares the results with existing state of the art in the field. Also, considering the fact that we can not separate just the encoder part of the model, we can not just compare the feature encoding capability of respective models, so we decided it would be better if we can just compare the output of the model, and the ability to reconstruct images. So we've compared the output of images by two models, and calculated root mean square value (RMSE) value for the constructed images.

### 4.1 Experimental Settings

- **Benchmark Dataset:** In this experimental work, we used primary dataset as, Multi-PIE [6] and 300wLP [18]. These datasets are labelled datasets collected in an artificially constructed environment. The dataset consists of 250 subjects, with 9 poses within  $\pm 60^\circ$ , two expressions and twenty illuminations. For the training purpose, we choose the first 200 for training and the remaining 50 for test. 300wLP contains view labels that are used to extract images with yaw angles from  $-60^\circ$  to  $+60^\circ$ , dividing them into 9 intervals. So, they can synchronize with Multi-PIE dataset after feeding into the model.
- **Implementation Details:** The network-implementation is modified from CR-GAN, where each of **E** shares the dual-pathway architecture with the **G**. The main structure for our model is adopted from the res-net (residual networks) as proposed in WGAN-GP [7], where E shares a similar network structure with D. During training **v** is set to 9 dimensional one-hot vectors where  $\mathbf{z} \in [-1, 1]^{19}$  in the latent embedding space. The batch-size we chose for our model is 20. We used Adam optimizer [11] with the learning rate of 0.0001 and momentum of [0.01, 0.89]. Choosing rest of the parameters of CR-GAN as default, we have  $\mathbf{C}_1 = 10$ ,  $\mathbf{C}_2 - \mathbf{C}_4 = 1$ , and  $\mathbf{C}_5 = 0.01$ . Finally, we train the model for 50 epochs.

### 4.2 Results and Discussion

The primary aim of the proposed model - DUO-GAN is to learn complete representation by using its dual-encoder architecture and dual-pathway architecture to span entire embedding space. We conduct experiments to evaluate these contributions with respect to CR-GAN. The comparative results are shown in Table 1. We can see how the model performs in the wild settings in Fig. 2.

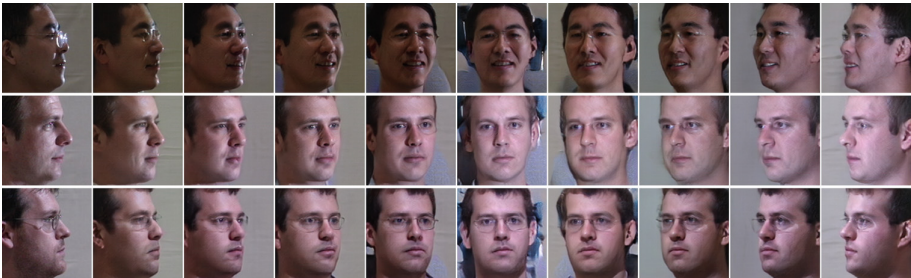
**Table 1.** Average RMSE(in mm) using dual encoder architecture, validated against CR-GAN.

Values	DUO-GAN	CR-GAN
Female subject	2.342 ( $\pm 0.501$ )	2.64 ( $\pm 0.491$ )
Male subject	2.4757 ( $\pm 0.143$ )	2.795 ( $\pm 0.52$ )



**Fig. 2.** Sample output on test images

In order to demonstrate the applicability of the proposed model, we compare it with four GANs, namely, BiGAN, DR-GAN, TP-GAN, and CR-GAN as depicted in Fig. 1. **CR-GAN** [14] used a dual-architecture for spanning embedding space, and learning better representation. Authors used a second reconstruction-pathway in order to make the encoder inverse of the generator. However, in practice, the Encoder doesn't seem to be powerful enough to span the entire embedding space. Comparatively, DUO-GAN uses two encoders in order to span the entire embedding space, which learns the representation comparatively more efficiently. The output produced by the proposed model is presented in Fig. 3.



**Fig. 3.** Sample output on similar, but unseen images

**DR-GAN** [15] also tackled the problem of generating multi-view images from a single image, through a linear network. Like in a linear network, input of decoder is the output of encoder, the model is not very robust to images outside the dataset. Comparatively, we use a second generation path, which leads to better learning and generalization.

**TP-GAN** [10] also used a dual-architecture for solving this problem. However, unlike our model, it uses two separate structure, i.e. these two structures don't share parameter, unlike our architecture. Further, these two independent architectures in TP-GAN aims to learn different set of features. Where as our architecture aims to learn collectively.

**Bi-GAN** [4] aims to learn collectively a **G** and an **E**. Theoretically, **E** should be an inverse of **G**. Because of their linear network, Bi-GAN leads to poor learning and doesn't lead to good generation especially for unseen data.

## 5 Conclusion

In this paper, we investigated different models and compared them for constructing multi-facet images from a single image. We propose a dual architecture model called DUO-GAN, which uses double duo-pathway framework for better learning the representation. The proposed model leverages the architecture to span latent embedding space in a better way and produces higher quality images in comparison to existing models.

**Acknowledgement.** I would like to express my special thanks of gratitude to my friend, Mr. Shivam Prasad who helped me in doing a lot in finalizing this paper within the limited time frame.

## References

1. Alqahtani, H., Kavakli-Thorne, M.: Adversarial disentanglement using latent classifier for pose-independent representation. In: International Conference on Image Analysis and Processing (ICIAP) (2019)
2. Alqahtani, H., Kavakli-Thorne, M., Kumar, G.: An analysis of evaluation metrics of gans. In: International Conference on Information Technology and Applications (ICITA) (2019)
3. Alqahtani, H., Kavakli-Thorne, M., Liu, C.Z.: An introduction to person re-identification with generative adversarial networks. arXiv preprint [arXiv:1904.05992](https://arxiv.org/abs/1904.05992) (2019)
4. Dumoulin, V., et al.: Adversarially learned inference. arXiv preprint [arXiv:1606.00704](https://arxiv.org/abs/1606.00704) (2016)
5. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
6. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. *Image Vis. Comput.* **28**(5), 807–813 (2010)
7. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems, pp. 5767–5777 (2017)
8. Hassner, T., Harel, S., Paz, E., Enbar, R.: Effective face frontalization in unconstrained images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4295–4304 (2015)
9. Hinton, G.E., Krizhevsky, A., Wang, S.D.: Transforming auto-encoders. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (eds.) ICANN 2011. LNCS, vol. 6791, pp. 44–51. Springer, Heidelberg (2011). [https://doi.org/10.1007/978-3-642-21735-7\\_6](https://doi.org/10.1007/978-3-642-21735-7_6)
10. Huang, R., Zhang, S., Li, T., He, R.: Beyond face rotation: global and local perception gan for photorealistic and identity preserving frontal view synthesis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2439–2448 (2017)



11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
12. Li, Y., Yang, M., Zhang, Z.: Multi-view representation learning: a survey from shallow methods to deep methods. arXiv preprint [arXiv:1610.01206](https://arxiv.org/abs/1610.01206) (2016)
13. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 397–403 (2013)
14. Tian, Y., Peng, X., Zhao, L., Zhang, S., Metaxas, D.N.: Cr-gan: learning complete representations for multi-view generation. arXiv preprint [arXiv:1806.11191](https://arxiv.org/abs/1806.11191) (2018)
15. Tran, L., Yin, X., Liu, X.: Disentangled representation learning gan for pose-invariant face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1415–1424 (2017)
16. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: learning single-view 3d object reconstruction without 3d supervision. In: Advances in Neural Information Processing Systems, pp. 1696–1704 (2016)
17. Yang, H., Mou, W., Zhang, Y., Patras, I., Gunes, H., Robinson, P.: Face alignment assisted by head pose estimation. arXiv preprint [arXiv:1507.03148](https://arxiv.org/abs/1507.03148) (2015)
18. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: a 3d solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 146–155 (2016)
19. Zhu, Z., Luo, P., Wang, X., Tang, X.: Multi-view perceptron: a deep model for learning face identity and view representations. In: Advances in Neural Information Processing Systems, pp. 217–225 (2014)