Extending the Capacity of CVAE for Face Synthesis and Modeling

Shengju Qian
UESTC
qianshengju@std.uestc.edu.cn

Wayne Wu
Tsinghua University
wwy15@mails.tsinghua.edu.cn

Yangxiaokang Liu
UESTC
liuyangxiaokang@std.uestc.edu.cn

Beier Zhu
Tsinghua University
zbe16@mails.tsinghua.edu.cn

Fumin Shen
UESTC
fumin.shen@gmail.com

Abstract

In this work, we develop an equilibrium conditional variational autoencoder(CVAE) for structure-guided face synthesis. The proposed model has a flexible appearance encoding space according to each facial geometry characteristic. Traditional CVAE, constrained by its capacity, tends to lose diversity and details on generating faces in unseen poses, resulting in structure incoherence or beautified phenomenon. By extending the capacity of CVAE in an unsupervised fashion, our model can thus generate photo-realistic unseen faces conditioned on a different structure, as well as disentangling appearance and structure features. Furthermore, with great disentanglement and generalization property, our model is able to deal with extreme expression and rotation.

1 Introduction

Conditional face synthesis requires the ability to understand and respond to images. Just like a baby, the model needs to gradually learn to be aware of a person's facial expression and geometry structure. This task, in other words, requires the power of disentangling appearance and structure. By assembling internal appearance and external geometry structure of different people, the model would be able to preserve given information and synthesize photo-realistic images.

Despite impressive results shown by GAN based variants[1, 2], modeling both facial geometry and appearance and interplaying between these two factors when generating images always needs multiple movements of a single instance to exist in training image series. As addressed, these tasks can be decomposed as disentangling facial shapes and appearance feature. Thus we propose a 'two-branch' conditional autoencoder which captures potential structures of a single face by utilizing its easily available landmarks. Conditioned on the latent representation of appearance, the model generates new images via the mapping from a target shape. There have been plenty of trails of disentangling [3] and image synthesis [4] using VAE or CVAE. Such models are good at modeling spatial transforms and naturally generating target images.

However, a common observation is that synthesized faces from these models lack diversity and tend to be a 'mean' face, which means that the estimated posterior distribution given the image is not accurate and distinguishable enough. Using traditional CVAE[5, 4, 6] equipped with a fixed

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

Gaussian prior, the learned conditional posteriors tend to collapse to a single mode which leads to the synthesized image's lack of details , especially training on outdoor datasets where identity variances are larger. To improve the behavior of CVAE, we propose a novel approach using K Gaussian priors in latent space z with different means and standard deviations, corresponding to different structural characteristics. Specifically, we identify each mode as a cluster center by clustering landmarks into K clusters. In our settings, each cluster tends to represent a facial characteristic, such as a side face, closed eyes. Due to the intractability of Gaussian mixtures in VAE, we also propose a more flexible prior which combines multiple gaussians into one equilibrium gaussian prior. If a given face contains multiple characteristics such as a side-face laughing person with a wide-open mouth, our formulation can better preserve these properties as well as providing better appearance details.

2 Problem Formulation

Face synthesis conditioned on appearance and pose code such as head pose, facial expressions relied on the sturdy power of robust representations learned by generative models. For instance, face rotation and frontalization[7, 8] series learn geometry and texture feature; for face expression animation[2] and face editing[9], related priors such as action units(AU) or 3DMM[10] are utilized.

Let x to denote an input RGB image from dataset X. For all $x \in X$, overall face is assumed to be driven by two latent factors, an appearance factor denoted as z which corresponds to identity, lighting and other so-called invariant properties of a single face, and a shape factor denoted as c which corresponds to variants including facial expression, head pose and so on. Our aim is: given an image of x and its estimated shape $\hat{y}$, a representation of appearance z needs to be learned, where we should be able to capture arbitrary faces given possible $\hat{c}$ with appearance across z. To this end, our image generator can be modeled to maximize $p(x|c, z)$, conditioned on c and z. For the generator itself, being able to generate another view (shape) of any given person and generate any person at a fixed given view (shape) is required.

3 Methodology

Fig. 1 is an overview of our method, the clusters are clusters of facial geometry structure, represented by sketches through interpolation between landmarks.

Conditional Variational Autoencoder In order to model the conditional distribution of $p(x|c)$, where x is the desired image and c is some representation of pose code of the input image. Thus, the VAE power can be straightforwardly extended by conditioning the desired distribution on c: estimated shape given. More specifically, by utilizing the conditional prior, potential semantic correlations and pixel affinity are likely to be captured. For instance, a person who is laughing is likely to have his/her eyes closed. Intuitively, by maximizing the lower bound on the conditional data-log-likelihood $p(x|c)$, we are able to train the encoder and decoder using gradient descent.

$$\log p(x|\hat{c}) = \log \int_z p(x, z|\hat{c}) dz \geq \mathbb{E}_q \log \frac{p(x, z|\hat{c})}{q(z|x, \hat{c})} = \mathbb{E}_q \log \frac{p(x|\hat{c}, z)p(z|\hat{c})}{q(z|x, \hat{c})} \tag{1}$$

In practice, following objective is typically used by representing the parameters for the encoder and decoder using θ,ϕ respectively:

$$\max_{\theta,\phi} \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(x^i|z^i, c^i) - D_{KL}[q_\phi(z|x, c), p(z|c)], s.t. \forall i, z^i \sim q_\phi(z|x, c) \tag{2}$$

Following reparametrization trick, the networks can be trained end-to-end using standard gradient descent, by restricting the encoder distribution $q_\phi(z, c)$ to be a zero-mean and standard-deviation Gaussian.

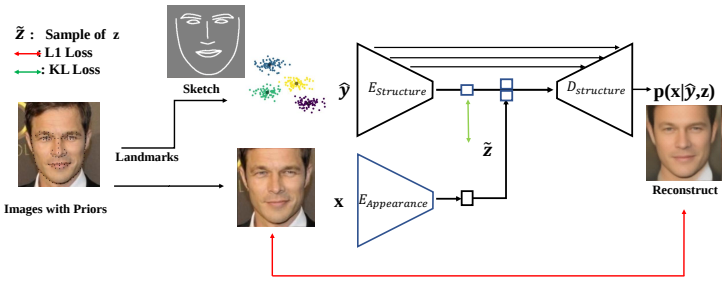


Figure 1: Overview of our method. In our setting, the priors are represented by landmarks.We used a open-sourced 98-point landmarks detector in [11]. Not that $\hat{y}$ in figure refer to c due to its ambiguity

3.1 Extending Capacity of Conditional VAE

As the stochastic objective of CVAE typically approximates the expectation using samples from the approximate posterior. One observation of 'Vanilla' CVAE trained on face feature disentangling is that its heavy dependence on the choice on the prior $p(x|c)$, which structures the learned appearance latent space via KL-divergence Loss term in Eq. (2). While the choice of fixing the prior to a zero-mean unit-variance Gaussian is empirical and computationally convenient, thus generated faces have less diverse and a smoothed face. This phenomenon appears frequently when the model is trained on outdoor faces series where sharp distinctness between different identities in terms of lighting, skin, background and so on is present. A key factor is that given a simple distribution, some peculiar identity features might be viewed as outliers. Clearly, the fixed form of prior brings about less diversity and inadequate descriptive power. Thus, the prior should also be changed over the external characteristics of a given face while being able to compute the KL divergence in a closed form. Straightforwardly, by encouraging the latent z space to be composed of K clusters, each corresponding to a different facial geometry characteristic, we can increase the representational power. Given an image I, we assume that we can obtain a distribution with k modes denoted by $c(I) = (c_1(I), c_2(I), ..., c_k(I))$. In our work, we identify these modes with a set of different views by clustering face sketches. Note that this formulation is general and can be applied to other definitions of a cluster, including using face recognition networks or in an unsupervised fashion[12].

Gaussian Mixture Latent Space: By modeling $p(z|c)$ as a Gaussian Mixture, for each cluster k with weight c_k, means μ_k and standard deviation σ_k:

$$p(z|c) = \sum_{k=1}^{K} c_k \mathcal{N}(z|\mu_k, \sigma_k^2 I) \tag{3}$$

As this form is not directly tractable to optimize Eq.(2) with GMM encoding, we thus need to approximate the KL divergence term stochastically by drawing the cluster component according to its cluster probability and then sample z from the selected Gaussian component. However, with the sampling step, at test time we also need to sample a component and its corresponding distribution. A limitation remains that using a defined clustering standard, if a face contains more than one cluster's characteristic, e.g. laughing and side face, it would still be conditioned on a single characteristic.

Equilibrium Gaussian Latent Space: In order to structure the z space to be accurate and concurrent. Motivated by [13], we propose a simple extended CVAE with an additive equilibrium Gaussian prior. Given k representative face clusters, each referring to one major facial structure characteristic. If a face image contains several characteristics with weights c_k, defined by a normalized cosine similarity with respective cluster

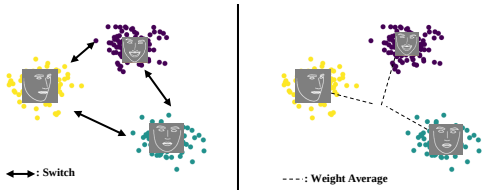


Figure 2: Illustration of difference between Gaussian Mixture and Equilibrum Gaussian latent space mode.

center, each corresponding to its mean μ_k in z latent space. $p(z|c)$ can be formulated as

$$p(z|c) = \mathcal{N}(z|\sum_{k=1}^{K} c_k\mu_k, \sigma^2 I) \tag{4}$$

where $\sigma^2 I$ is a covariance matrix with $\sigma^2 = \sum_{k=1}^{K} c_k^2\sigma_k^2$. The difference is shown in Fig 2. With the extended Gaussian space in CVAE model, the KL-divergence should be computed over $q(z|x,c) = N(z|\mu(x,c), \sigma^2(x,c)I)$. The KL-divergence can be derived to be:

$$D_{KL} = \log(\frac{\sigma}{\sigma_\phi}) + \frac{1}{2\sigma^2}E_{q_\phi}[(z - \sum_{k=1}^{K} c_k\mu_k)^2] - \frac{1}{2} = \log(\frac{\sigma}{\sigma_\phi}) + \frac{\sigma_\phi^2 + (\mu_p hi - \sum_{k=1}^{K} c_k\mu_k)^2}{2\sigma^2} - \frac{1}{2} \tag{5}$$

For each cluster, we keep it fixed by randomly sampling means and same constant deviations for simplicity.

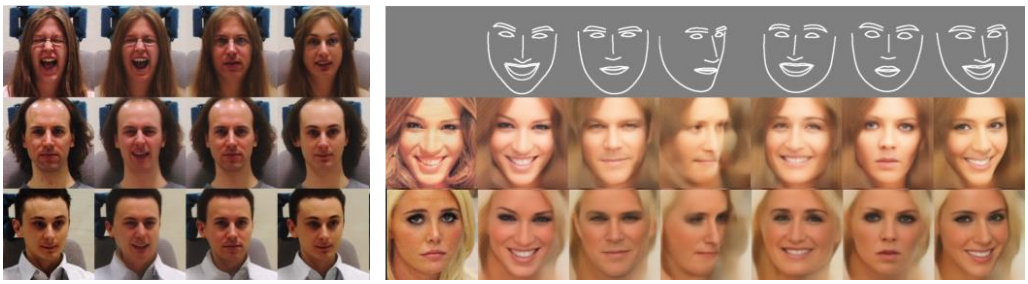
4 Experiments

We conduct experiments and train our models on MultiPIE [14], CelebA [15]. Landmarks are obtained using the pre-trained network from [11]. We first center crop the image to 256×256 for preprocessing. For the network itself, we use U-Net[16] for structure encoding and decoding, where spatial information can be better preserved by skip connection. For training procedure, Layer-sequential unit-variance initialization [17] and Adam optimizer [18] is used.

When testing, we are able to generate multiple poses and expressions represented by landmarks of a single person. And the proposed model is evaluated on two tasks including:

Emotion Transfer By fixing one's appearance property, transferring another person's expression onto a given face.

Conditional Pose Generation When given a single image, our model is expected to generate itself under arbitrary pose and structure where appearance such as identity, lighting, and background are preserved. We randomly choose three people, generating his/her faces by giving different structures from themselves. The result is shown in Fig. 3



(a) Emotion Transfer

(b) Conditional Image Generation

Figure 3: Results on MultiPIE and CelebA dataset. (a) Emotion transfer: leftmost row is are source images, next three rows are generated faces where each person gets its emotion transferred driven another's (b) Conditional Pose Generation: Given 6 target poses, generate faces under target poses.

5 Conclusion

We pose that the problem of synthesized image's quality depends heavily on the capacity of existing encoding space. By modeling so-called "z" space with an equilibrium Gaussian latent space, results on face synthesis show that our approach can well preserve internal complex appearance using a simple CVAE structure, thus being able to generate photo-realistic faces with diversity.

Further Work We would like to investigate more general formulations about the condition c, not just using a simple structure clustering. They can be obtained using other visual features. Also, relations between different clusters is also an interesting topic to investigate.

References

[1] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: Fine-grained image generation through asymmetric training. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2764–2773. IEEE, 2017.

[2] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2018.

[3] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[4] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018.

[5] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[7] Rui Huang, Shu Zhang, Tianyu Li, Ran He, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *arXiv preprint arXiv:1704.04086*, 2017.

[8] Yibo Hu, Xiang Wu, Bing Yu, Ran He, and Zhenan Sun. Pose-guided photorealistic face rotation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018.

[9] Linsen Song, Jie Cao, Linxiao Song, Yibo Hu, and Ran He. Geometry-aware face completion and editing. *arXiv preprint arXiv:1809.02967*, 2018.

[10] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2018.

[11] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018.

[12] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. *arXiv preprint arXiv:1807.05520*, 2018.

[13] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems*, pages 5756–5766, 2017.

[14] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[17] Dmytro Mishkin and Jiri Matas. All you need is a good init. *arXiv preprint arXiv:1511.06422*, 2015.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.