Neural Image Completion Based on Label Differentiation

Xianyi Zhu¹, Zixiao Ye², Guanghua Tan²*, Sheng Duan¹, Yaohui Liu¹, Dong Liu¹

¹College of Software and Communication Engineering, Xiangnan University, Chenzhou 423000, China

²College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China

*ghtan81@foxmail.com

Abstract. Deep learning based image completion methods are generally based on three technologies, namely the auto encoder based method, the generation of adversarial networks based method and the recurrent network based method. However, many methods have very single results. In order to obtain the diversity of the completion results, this paper proposes a neural image completion method based on label differentiation, called LD-PICNet (Label Differentiation PICNet). This method can not only generate a complete image with clear and good semantic information, but also actively edit the tags on the generated results to maximize the diversity of the output. Specifically, this paper introduces an auxiliary classifier, which uses a single label of the image ground truth to reconstruct the image, and uses the label to actively increase the difference of the latent vectors during image completion to achieve variability of output. In addition, this paper also introduces a depth-weighted loss function based on information entropy. Different data sets are used to conduct experiments. The model’s ability to complete different types of targets is tested. The results show that this method has the ability to generate diversified results, and it has higher clarity than the other methods.

Keywords: Image Completion, Label Differentiation, Deep Learning.

1. Introduction

Image completion technology refers to repairing a broken picture by computer method. Image completion is a problem between image editing and image generation. Image completion was originally a problem of traditional graphics. The purpose of image completion is to let the computer estimate the missing information in the image according to the designed rules, and restore the image to the greatest extent possible. Its essence is to reconstruct the complete information from the known incomplete information, and high-quality image completion Not only the generated content has reasonable semantics, but also the generated image texture is clear and true enough. This type of problem has important theoretical research significance in the field of artificial intelligence (see Fig. 1).

There are many implementation methods for image completion technology. The output of most methods is very simple, and only one completion result can be generated for each defective image input. In the classic learning-based method, there is usually only one ground truth training instance per label, which will make the completion result singular, even if sampling from the variational auto
encoder (VAE) [1], it will still face insufficient diversity problem. Because the probability space corresponding to the possible result of each defective image is very large.

In order to enhance the output diversity of the image completion method, this paper proposes a new loss function based on the difference of appearance matching-depth-weighted loss function. The loss function gives different weights from the edge to the interior of the masked image, enhances the generalization ability of the model to the depth of the masked area, and makes the model show greater differences in the generated results. At the same time, we improve PICNet[2], adding a tag-led difference control module, combined with labeling information, to enhance the controllability of the completion direction, and enhance model clarity. Experiment results show that our model can produce diverse and highly realistic image completion results.

Fig.1 From left to right are: a masked image, an original image, and our results (the last three)

2. Our Network Model

The method in this article is based on PICNet, but there are many important changes. Our network architecture is shown in Fig.2, using two parallel path frameworks based on the principle of probability: the upper part of the path is the reconstruction path, using the given data set ground truth to get the missing part prior, and reconstruct the original from the distribution image; the next part of the path is the generated path, coupling the condition prior and the distribution in the reconstructed path, both of which are based on GAN[3].

The upper and lower paths of LD-PICNet are composed of two similar variational auto encoders (VAE)[1]. The encoder encodes the input image to obtain the latent space of the image, and samples the latent space to obtain noise samples. The information is fused, and then the reconstruction and generation tasks are performed. At the same time, the discriminator not only needs to judge the authenticity of each sample, but also needs to give the label of the image. Through the control of the label vector to achieve the intervention of the generated results.

Fig.2 LD-PICNet network frame diagram

Label differentiation. A difference control module is proposed in this method, in order to expect the model to obtain the ability to directionally acquire the diversity of the image completion result, and the rich auxiliary information can also enhance the completion image the quality of. Since PICNet is uncontrollable when sampling the latent space of the missing image, it cannot fully cover the entire
probability space, so that the diversity of the generated results still has room for improvement (such as controlling the expression of the facial image completion).

As shown in Fig.2, when the latent space of the image generated by the encoder is sampled, the tag information $c_i$ and the latent vector $z_c$ are fused, and the tag information is used to guide the decoder of the reconstruction path and the generation path. The discriminator is a deep convolutional neural network. The discriminator's loss function is changed to consist of two parts, which are the judgment of true and false samples and the prediction of sample labels:

$$\mathcal{L}_D = \alpha_S \mathcal{L}_S + \alpha_L \mathcal{L}_L$$

$\mathcal{L}_S$ is the discriminant loss of the input true and false samples, which is the log likelihood of positive samples; and $\mathcal{L}_L$ is the label prediction loss, which is the log likelihood of the correct label.

Depth-weighted loss function. Compared with the PICNet, a new depth-weighted loss function is introduced in our method, which enhances the model's ability to generalize image gaps. In the depth of the missing part, the entropy value to be reduced should be more than the marginal area, that is, the degree of freedom left to the model for the completion task should be higher, and the "imagination" of the model should be more abundant. The probability of a specific pixel's value increases as it goes deeper into the missing area. The higher the information entropy of this part, the smaller the probability of taking a certain pixel. Thus, we propose to solve a weight coefficient for the missing areas of the image. The appearance matching loss functions $\mathcal{L}_{app}^r$ and $\mathcal{L}_{app}^g$ add of the model can be rewritten as:

$$\mathcal{L}_{app}^r(i,j) = \parallel \overset{\wedge}{M}w(i,j) \odot (l_{rec}(i,j) - l_{g}(i,j)) \parallel_1$$

$$\mathcal{L}_{app}^g(i,j) = \parallel \overset{\wedge}{M} \odot (l_{gen}(i,j) - l_{g}(i,j)) \parallel_1$$

where mask $M_{w}(i,j) = \frac{-\text{sigmoid}(\text{layer}(i,j)\times0.75-5))}{2} + 1$ and mask $\overset{\wedge}{M}(i,j) = \frac{1}{\exp(\text{layer}(i,j))}$.

Overall loss function. The overall loss function is composed of three modules (see Eq. (4)). The first group ($\mathcal{L}_{\text{KL}}^r + \mathcal{L}_{\text{KL}}^g$) is the distribution regularization loss, using the VAE framework, and the KL divergence is used to standardize the consistency between the distribution pairs. The second group ($\mathcal{L}_{app}^r + \mathcal{L}_{app}^g$) is the appearance matching loss, used to restore the unmasked part. The third group ($\mathcal{L}_{\text{ad}}^r + \mathcal{L}_{\text{ad}}^g$) is against losses, making the output more in line with the distribution of the training set. Each group of losses has two parts, which are from the reconstruction path and the generation path.

$$\mathcal{L} = \alpha_{\text{KL}}(\mathcal{L}_{\text{KL}}^r + \mathcal{L}_{\text{KL}}^g) + \alpha_{\text{app}}(\mathcal{L}_{app}^r + \mathcal{L}_{app}^g) + \alpha_{\text{ad}}(\mathcal{L}_{\text{ad}}^r + \mathcal{L}_{\text{ad}}^g)$$

3. Evaluation and analysis

In order to verify the effect of the image completion method in this article on different targets, we used CelebA [5] face dataset, Paris [6] Paris street view dataset, and ImageNet [7]. Some categories of pictures to experiment. Since the image completion task does not have a unique solution, and in order to present the diversity of the output results, 50 different results are generated for each picture to be completed, and the 10 highest scores of the discriminator are selected as their effect display and tested. Various masking effects for missing areas have been added. The content of each data set is roughly as shown in Fig.3.
In the experiment, all the pre-trained models used are public. The model contains nearly 6M trainable parameters. In the loss function, $\alpha_{\text{KL}}$ is initialized to 20, $\alpha_{\text{app}}$ is initialized to 20, $\alpha_{\text{ad}}$ is initialized to 1. The initial learning rate is set to $\lambda = 10^{-4}$. All pictures are preprocessed to $256 \times 256$ pixels. The masks in the training process are mainly divided into three categories, the central $128 \times 128$ square mask, the random square mask and the random irregular mask, roughly as shown in Fig.4.

Qualitative comparison. In order to obtain a qualitative index, this paper only uses the discriminator to give the output results of the 10 completion tasks with the highest confidence, and selects a sample with the best visual effect for comparison. In comparison, 20,000 images in the ImageNet test set were used for the experiment, and four qualitative indicators, such as average L1 loss, peak signal-to-noise ratio (PSNR), total variation (TV), and Inception Score [8] (IS) were used for comparison. The result as shown in the Tab.1.

| Methods     | L1 loss | PSNR  | TV loss | IS    |
|-------------|---------|-------|---------|-------|
| GL [15]     | 15.32   | 19.36 | 13.97   | 24.31 |
| CA [16]     | 13.57   | 19.22 | 19.55   | 28.80 |
| PICNet      | 12.91   | 20.10 | 12.18   | 24.90 |
| LD-PICNet   | 12.67   | 20.31 | 12.67   | 25.11 |

**Tab.1** Qualitative comparison of models trained with the central square missing mask

**Fig.3** The dataset examples

**Fig.4** The image masks

**Fig.5** Completion comparison of face images with square masks in the center, from left to right are (a) Input, (b) Ground truth, (c) Shift-Net (SN), (d) Contextual Attention (CA) and (e) this method
Fig. 6 Completion and comparison of the center square masks street image, from left to right are (a) Input, (b) Ground truth, (c) PM, (d) Context-Encoder (CE) [4], (e) Shift-Net (SN) and (f) this method.

Fig. 7 Completion and contrast of the central square masks landscape image, from left to right are (a) Input, (b) Ground truth, (c) PM, (d) CE, (e) CA, and (f) this method.

Quantitative comparison. For the face data set, the experiment uses 200,000 cropped and aligned face images, and uses the first effective attribute label (such as blond hair, thick eyebrows, etc.) for each image for training. As shown in Fig. 5, the face images restored by Contextual Attention [9] (CA) and ShiftNet [10] (SN) have strong semantic information and satisfactory clarity, but can only generate one for each completion results, some poor output cannot be changed. The experiment shows that when using different tags for differential image completion, because the known part of the image cannot be changed, although it cannot achieve the huge difference like image generation [11], it has been able to make the model generation as much as possible.

See Fig. 6 and Fig. 7. For the image completion tasks of street scenes and landscape pictures, the Paris dataset is used for training. It can be seen that the traditional method based on sample matching [12] (PatchMatch, PM) works well in the face of reproducible repetitive structures inside the image, but cannot learn high-level semantic information. As Context-Encoder (CE), the semantic information has a relatively clear completion result, but for each input, only a fixed output can be generated, and its resolution is not high. SN has a better effect on structural information completion, but it will tend to blur. The method in this paper can not only obtain high-level semantic information, maintain structure propagation, but also take into account the underlying texture, and the clarity is also higher.

4. Conclusion
This paper proposes a neural image completion method, LD-PICNet. Experimental results on multiple different types of datasets have achieved good performance. Compared with basic PICNet, our LD-PICNet includes the difference control module and the depth-weighted loss function that enhances the output diversity of the model are introduced. Experiment results verify the effect of the difference control module and the depth-weighted loss function, and compare with other methods.

Acknowledgments
This work was supported by the Natural Science Foundation of Hunan Province (No. 2018JJ3479) and Scientific Research Fund of Hunan Provincial Education (No.18B504).
References

[1] Kingma D P and Welling M 2014 2nd Int. Conf. on Learning Representations, ICLR 2014, Ban, AB, Canada, April 14-16, 2014, Conf. Track Proc. ed Bengio Y and LeCun Y URL

[2] Zheng C, Cham T and Cai J 2019 IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019 (IEEE Computer Society) pp1438-1447

[3] Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A C and Bengio Y 2014 Advances in Neural Information Processing Systems 27: Annual Conf. on Neural Information Processing Systems 2014, December 8-13 2014 Montreal Quebec, Canada ed Ghahramani Z Welling M et al pp2672-2680

[4] Pathak D, Kr‘ahenb‘uhl P, Donahue J, Darrell T and Efros A 2016 A 2016 IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2016 Las Vegas NV USA June 27-30 2016 (IEEE Computer Society) pp2536–2544

[5] Liu Z, Luo P, Wang X and Tang X 2015 2015 IEEE Int. Conf. on Computer Vision, ICCV 2015, Santiago Chile December 7-13 2015 (IEEE Computer Society) pp3730–3738

[6] Doersch C, Singh S, Gupta A, Sivic J and Efros A A 2015 Commun. ACM 58 pp103–110

[7] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M S, Berg A C and Li F 2015 Int. J. Comput. Vis 115 pp211–252

[8] Barratt S T and Sharma R 2018 Preprint 1801.01973 URL

[9] Yu J, Lin Z, Yang J, Shen X, Lu X and Huang T S 2018 2018 IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22 2018 (IEEE Computer Society) pp5505–5514

[10] Yan Z, Li X, Li M, Zuo W and Shan S 2018 Computer Vision - ECCV 2018 - 15th European Conf., Munich, Germany, September 8-14, 2018, Proc., Part XIV (Lecture Notes in Computer Science vol 11218) ed Ferrari V Hebert M et al (Springer) pp3–19

[11] Oord A, Kalchbrenner N, Espeholt L, Kavukcuoglu K, Vinyals O and Graves A 2016 Advances in Neural Information Processing Systems 29: Annual Conf. on Neural Information Processing Systems 2016 December 5-10 2016 Barcelona Spain ed Lee D D Sugiyama M et al pp4790–4798

[12] Barnes C, Shechtman E, Finkelstein A and Goldman D B 2009 ACM Trans. Graph 28 p24