

Research on Cross-modal Image Generation Based on Generative Adversarial Network

Lujuan Zhang

Guangdong University of Science and Technology, Guangdong, China

Abstract

This paper proposes an cross-modal image generation network, GAN2C, by imposing dual consistency constraints (2C) to a closed loop encoder-decoder architecture based on the generative adversarial nets (GAN). When adopting deep neural networks as function approximators, GAN2C enables highly effective multimodality image generation with sparse observation in the target modes. For empirical demonstration and model evaluation, we show that trained deep neural networks in GAN2C can infer colors for grayscale images. The results of the experiment show that compared with the classic image-to-image conversion algorithm, the method in this paper can generate a color image that is close to the real one.

Keywords

Generative Adversarial Nets; Image generation; Cross-modal; Colorization.

1. Introduction

Multi-modal images are images that describe different observation information of the same object obtained by using different technical means. We consider the problem of generating a certain modal image of the scene, given other modal images of the same scene, and one modal of images are more expensive/difficult to obtain than another. For example, taking dense measurements of depth is generally more difficult than taking an RGB picture. In order to solve the problem of difficulty in acquiring a certain modal image, in this work, we propose a cross-modal image generation scheme, GAN2C.

The cross-modal image generation can use a certain modal image (source modal image) of the object to generate another modal image (target modal image). Specifically, the development in GAN2C is designated to where a small fraction of the target-mode image is observed. In other words, our goal is to use the source modal image and a small amount of sparse target observation information to generate the target modal image.

We apply the proposed model to the experiment of color restoration from grayscale images, and got excellent results. In addition, our proposed model can also be used to predict depth from RGB images.

2. Related Works

2.1. Image to Image Translation

In image processing based on deep learning, many problems can be considered as image-to-image translation tasks. Image-to-image translation refers to transforming an image from one representation scene to another representation scene, which involves the translation and generation of images. The recent development of deep neural networks [1] along with the GAN [2] framework provides a uniform solution to many related problems. For instance, Isola et al. [3] used a conditional generative adversarial network (cGAN) [4] to learn the mapping from conditional input image to its corresponding output image, which achieved good results in

various of image-to-image translation tasks. Recently, several proposed methods [5, 6] tackled the issue with unpaired image-to-image translation frameworks. Zhu et al. [5] proposed a general model that does not require paired training data—CycleGAN, which uses two GAN networks to form a ring network structure.

2.2. Colorization

Colorization of grayscale images has been a hot topic in computer vision [7, 8]. For example, Zhang et al. [9] take it as a classification task and aims to produce plausible colorizations that could potentially fool human. Hasnat et al. [10] proposed a grayscale image colorization method using a reference color image of the same type as the query image. Koushkaki et al. [11] considered the colorization as a process that deviates from the balance of grayscale and proposed a technique using tensor decomposition to extract and gather overall color information from a reference color image to guide the direction of deviation.

3. Method

3.1. Probabilistic Model

We are interested in the problem with following setting where there are two modes of information, X and Y . One can completely observe X and partially observe Y , and wants to estimate complete Y from the information of X and incomplete Y . To make the problem more explicit, we introduce a probabilistic graphical model, as illustrates in Figure 1. Formally, Z is the latent variable that represents the set of all the hidden information of the object, in which X and Y are two modes information of concern; $\tilde{Y}^s$ represents a subset of information in Y . The problem is to estimate Y from X and $\tilde{Y}^s$. In other words, our goal is to construct the probabilistic model, $P(Y|X, \tilde{Y}^s)$.

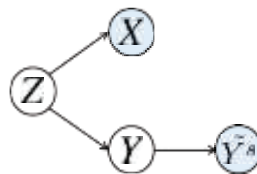


Figure 1. Probabilistic Model of Information Completion of Two Modes.

3.2. Network

Given the mode- X image and mode- Y measurements, the problem is to obtain the complete mode- Y image. To realize the probabilistic model with the structure illustrated in Figure 1, we adopt the architecture of GAN. Similarly, we use G as our $\mathcal{X} \rightarrow \mathcal{Y}$ map, and expect D implicitly imposing that the generated $G(X)$ follows the posterior probability $Y|X$. Specifically, in our information completion framework, in addition to the fully observed X , the generator G also takes as input the partially observed $\tilde{Y}^s$.

Moreover, in order to couple X and Y more tightly, we employ a variant of the GAN scheme, Cycle-consistent adversarial networks (CycleGANs). The CycleGANs introduce an extra pair of generator-discriminator to build the inverse map. On the one hand, for the generators, we apply the adversarial loss to induce the output distribution match the target distribution, and apply two consistency constraint losses, the reconstruction loss to ensure the generated image preserve the feature of the original image and the sparse supervision consistency loss to emphasize the consistency at partial observed points.

Adversarial loss encourages the generators to produce samples following the target distribution (unable to be distinguished by the discriminator),

$$\mathcal{L}_{ADV} = \mathbb{E}_{X, \tilde{Y}^s} \left[\log(1 - D^Y(G(X, \tilde{Y}^s))) \right] + \mathbb{E}_Y \left[\log(1 - D^X(F(Y))) \right] \quad (1)$$

Reconstruction loss encourages the generated image structurally resembles that of the input image,

$$\mathcal{L}_{REC} = \mathbb{E}_{X, \tilde{Y}^s} \left[\|X - F(G(X, \tilde{Y}^s))\|_1 \right] + \mathbb{E}_{Y, \tilde{Y}^s} \left[\|Y - G(F(Y), \tilde{Y}^s)\|_1 \right] \quad (2)$$

Note that we use ℓ_1 loss to improve training stability. The same applies to the consistency loss (8) below.

Sparse supervision consistency loss ensures a sparse subset of predicted Y is consistent with the available information about the target mode,

$$\mathcal{L}_{SSC} = \mathbb{E}_{X, \tilde{Y}^s} \left[\|G(X, \tilde{Y}^s)[\mathbf{J}] - \tilde{Y}^s\|_1 \right] \quad (3)$$

where $\mathbf{J}$ represents the subset indexes in the predicted target mode (in Y space) corresponding to the information in $\tilde{Y}^s$.

Finally, the objective function for generators is

$$\mathcal{L}_G = \mathcal{L}_{ADV} + \mathcal{L}_{REC} + \mathcal{L}_{SSC} \quad (4)$$

Correspondingly, the loss function of the discriminators DX and DY are

$$\mathcal{L}_{D^X} = \frac{1}{2} \left(\mathbb{E}_X [1 - D^X(X)] + \mathbb{E}_Y [D^X(F(Y))] \right) \quad (5)$$

$$\mathcal{L}_{D^Y} = \frac{1}{2} \left(\mathbb{E}_Y [1 - D^Y(Y)] + \mathbb{E}_{X, \tilde{Y}^s} [D^Y(G(X, \tilde{Y}^s))] \right) \quad (6)$$

We can optimize losses (4), (5) and (6) in alternating iterations to train a GAN2C model.

Furthermore, Figure 2 shows the structures of the deep neural networks implementing GAN2C. The first pair of generator-discriminator networks (X and DY) is responsible for learning the mapping relationship from the source modal image to the target modal image. Among them, in addition to receiving the source modal image x as input, the generator G also includes sparse real target observations $\tilde{Y}^s$, and outputs the corresponding target modal image $G(x, \tilde{Y}^s)$; The discriminator DY judges the authenticity of the target modal image, including the real target modal image y and the image $G(x, \tilde{Y}^s)$ generated by the generator G . The other pair of generator-discriminator networks (F and DX) is responsible for the opposite learning process, mapping the target modal image to the source modal image. Among them, the generator F receives the target modal image y as input and outputs the source modal image $F(y)$; the discriminator DX is used to judge the authenticity of the real source modal image x and the generated source modal image $F(y)$. The two groups of GAN networks form a cyclic structure, which can realize the reconstruction of the source modal image x and the target modal image y , namely $x \rightarrow G(x, \tilde{Y}^s) \rightarrow F(G(x, \tilde{Y}^s))$, $y \rightarrow F(y) \rightarrow G(F(y))$.

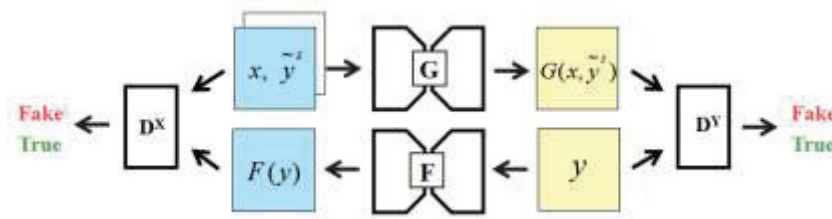


Figure 2. Structure of cross-modal image generation based on GAN2C

Generally, the network structures of generators and discriminators both use convolutional neural networks as basic components.

4. Experiment

This paper compares the proposed GAN2C with the classic image-to-image conversion algorithms Pix2Pix [12] and CycleGAN [5]. To test the versatility of our model, we collected 4000 natural landscape images from the Internet to train the model.

Figure 3 are the results of color images generated using different methods and sampling different amounts of color information. For example, "GAN2C-50" means that we sampled the true color information of 50 pixels for experimentation. It can be seen that GAN2C achieved superior performance compared to the CycleGAN, with additional partial observations. Remarkably, GAN2C also outperformed the direct predictive DNN in [12]. Moreover, there is almost no difference between an image generated using only color information of 50 pixels and a true color image.

This experiment proves that adding sparse target observations can greatly help the model to learn the one-to-one mapping relationship between the source modal image and the target modal image, thereby generating approximate real images, which is necessary for some images with strict and accurate output requirements. Image conversion tasks are essential.

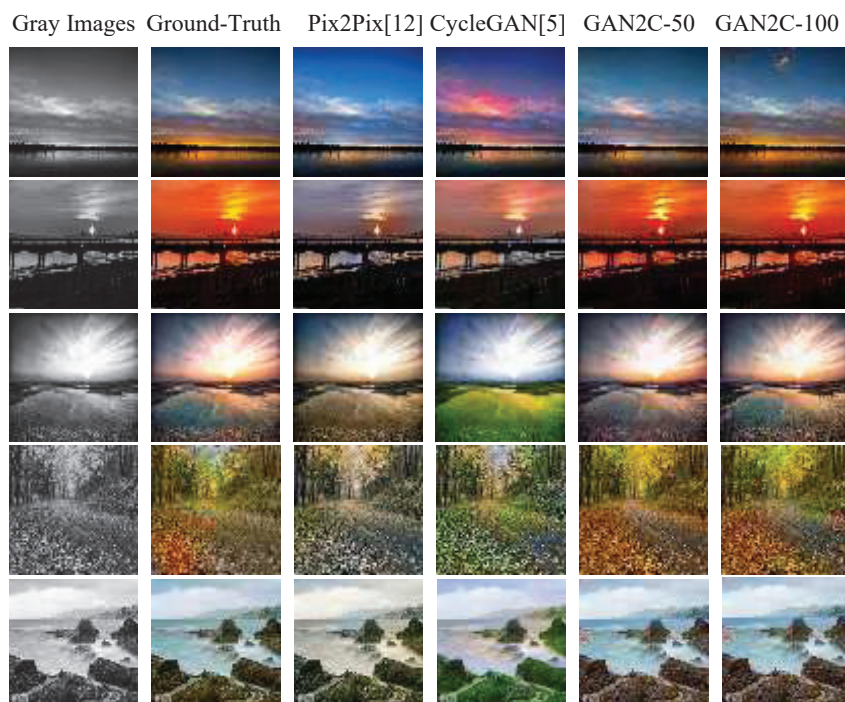


Figure 3. Image color restoration results using different methods

In order to verify the generalization ability of the image color restoration model based on GAN2C, this paper directly applies the model trained using the network image above to the DAVIS [13] video image data set to restore the color of the grayscale image. Figure 4 shows the effect of three different methods on the colorization of gray images on the DAVIS data set. The model used in our method is the model GAN2C-100 pre-trained in the previous experiment. Correspondingly, in the test, it is necessary to randomly sample the real color information of 100 pixels from the real color image. It can be seen from the figure that compared with the benchmark algorithm, the model trained using the method in this paper can generate color images that are closer to true colors.

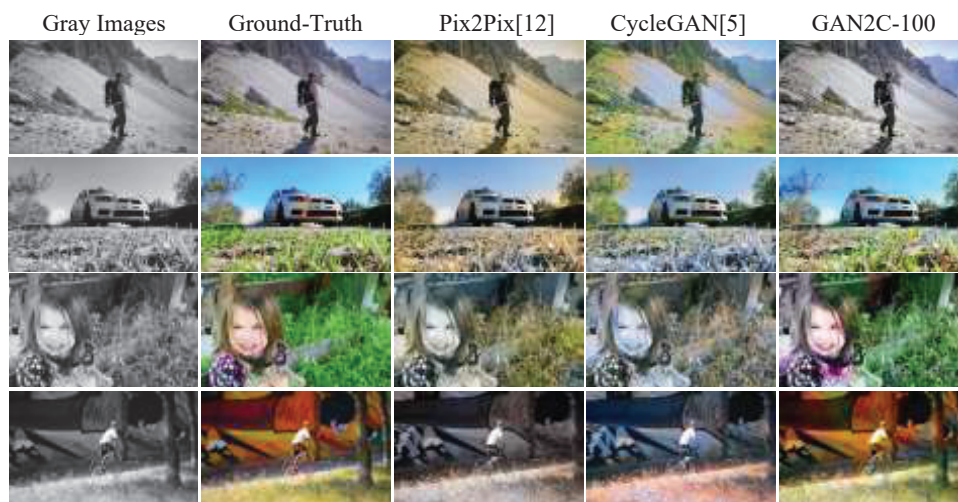


Figure 4. Image color restoration results of three different methods on the DAVIS dataset

5. Conclusion

In this paper, a cross-modal image generation model is proposed, and the performance of the model is verified by designing gray-scale image restoration experiments. In this experiment, this paper verifies the important effect of adding a small amount of real colors to the color restoration of grayscale images. At the same time, the pre-trained model using natural landscape images is applied to the color restoration of other data sets, which proves the transferability and generalization ability of the cross-modal image generation model.

The cross-modal image generation model can also be applied to some special scenarios, such as depth prediction. It is known that for a certain scene, we can easily use the camera to take a picture, and it takes a lot of manpower, material and financial resources to obtain the depth information of the corresponding scene, and due to the limitation of current technology, we can only reluctantly use radar equipment to scan to get a small amount of depth information. Therefore, we can regard the depth prediction problem as an image generation problem, and use the color image taken by the camera and a small amount of depth observation information to generate a complete depth picture using the GAN2C model proposed in this paper.

References

- [1] Y. LeCun, Y. Bengio, and G. Hinton: Deep learning, *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al: Generative adversarial nets, in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.
- [3] R. Zhang, P. Isola, and A. A. Efros: Colorful image colorization, in *European Conference on Computer Vision*, 2016, pp. 649–666.

- [4] D. Eigen, C. Puhrsch, and R. Fergus: Depth map prediction from a single image using a multi-scale deep network, in International Conference on Neural Information Processing Systems, 2014, pp. 2366–2374.
- [5] Zhu J Y, Park T, Isola P, et al: Unpaired image-to-image translation using cycle-consistent adversarial networks[C]. ICCV 2017: 2242-2251.
- [6] E. Shelhamer, J. Long: Fully convolutional networks for semantic segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640–651, 2017.
- [7] J. Li, P. Hao, and C. Zhang: Transferring colours to grayscale images by locally linear embedding,” in BMVC, 2008, pp. 1–10.
- [8] R. Zhang, J. Zhu, P. Isola, et al: Real-time user-guided image colorization with learned deep priors, arXiv preprint arXiv:1705.02999, 2017.
- [9] T. Welsh, M. Ashikhmin, and K. Mueller: Transferring color to greyscale images, vol. 21, no. 3, pp. 277–280, 2002.
- [10] H. R. Koushkaki, M. R. Salehi, and E. Abiri: Automatic colourization of grayscale images based on tensor decomposition, Multimedia Tools and Applications, pp. 1–21, 2017.
- [11] H. Wang, D. Huang, K. Jia: Hierarchical image segmentation ensemble for objectness in RGB-D images,” IEEE Transactions on Circuits and Systems for Video Technology, 2017.
- [12] Mirza M, Osindero S: Conditional Generative Adversarial Nets[J]. Computer Science, 2014:2672-2680.
- [13] <https://davischallenge.org/davis2017/code.html>