AIST4010 Project Milestone Report on Image Style Transfer

SHI Juluan

SID: 1155160208

1 Introduction

Since the advent of AlexNet in 2012, CNNs have become a dominant model in various computer vision tasks, including classification and segmentation. However, image style transfer poses a distinct challenge due to its subjective nature and the need to balance content and style effectively.

CNN and its variations have shown remarkable capacity for extracting hidden features from input images, enabling the synthesis of new images by combining features from different images. Nevertheless, to achieve satisfactory and robust results, careful model architecture design and loss function formulation are still needed.

For this project, I intend to delve into the history of the image style transfer task and provide an overview of the significant contributions made by various renowned papers. Ultimately, I aim to re-implement CocosNet[12], a formidable framework for image-to-image translation, and conduct a comparative analysis of its performance against other previously proposed models.

2 Related Work

One of the most intuitive methods for achieving image style transfer involves extracting the content from one image and the style from another, followed by combining them to produce a novel image. Gatys' group[3] successfully adopted this approach by utilizing the VGG model[10] for feature extraction and constructing a loss function that combines content loss and style loss for training. While the content representation can be easily extracted from each feature map layer, defining and quantifying style is a more challenging task. To address this issue, they used the Gram matrix to obtain the feature correlation (style representation) of the image. Despite the remarkable performance of this approach, the relationship between the Gram matrix and style representation remains unclear. In this regard, Yanghao Li[6] further proposed the idea that matching the Gram matrices of features maps is equivalent to minimize the Maximum Mean Discrepancy (MMD) with the second order polynomial kernel. That means neural style transfer is to match the feature distributions between the style images and the generated images.

While the above method for image style transfer have demonstrated outstanding performance and

a clear understanding of the underlying principles, manually designing loss functions for the feature maps is widely regarded as a challenging task. Moreover, above method are primarily suitable for creating artistic images that do not require a high level of detail. It can be challenging to design a loss function that converts daytime images to nighttime images while preserving the detailed features of the original images. In fact, Qifeng Chen et al[1] used a pretrained VGG to construct the loss function for generating high quality photographic images from semantic layouts and achieved satisfactory later on. However, the success of Chen's story also shows that the design of loss function is demanding.

Fortunately, with the advent of Generative Adversarial Networks (GANs)[2][4], solving some of these challenging tasks has become feasible and the solution is intuitive to understand. The discriminator in GANs can differentiate between labeled and generated fake images, thereby assisting the generator in producing higher quality results that are indistinguishable from labeled images. In this context, Phillip Isola and his team[5] proposed pix2pix using conditional GANs (cGANs) for more general image style transfer tasks. The loss function is composed of standard GAN loss and L1 loss. The GAN loss is used for learning new styles of the image while the L1 loss is used for preserving the spatial properties between the generated images and target images. Building on pix2pix, Ting-Chun Wang[11] used a multi-scale generator and discriminator architecture to generate style-transferred images with higher quality. While traditional image style transfer tasks have achieved satisfactory performance, generating high-quality photographic images from semantic layouts still has some potential challenges. Unconditional-based normalization methods may clear the semantic information during training, resulting in unrealistic generated results. To tackle this problem, Park and his team[9] proposed spatially-adaptive normalization (SPADE), which uses a set of learned affine parameters that are conditioned on the semantic label of the input image. These affine parameters are used to normalize the activations of the feature map in a way that is specific to the semantic content of the image.

Thanks to the efforts of many researchers, generating high-quality images from conditional inputs has become practical. In recent years, Rui Liu and his team[7] integrated contrasting learning[8] with existing conditional generative adversarial networks to produce diverse output images. It is clear that AI's performance in computer vision is becoming increasingly impressive.

In addition to the powerful supervised learning models mentioned above, CycleGAN[13] is also worth mentioning because it relies only on unpaired images for style transfer tasks. As the model's name suggests, it was trained in a cyclical manner using two pairs of generator and discriminator to learn the style and content of images. However, the cycle consistency loss constraint in the model tends to cause the output image to follow the shape of the input images, which can make it difficult to generate images that require changes in input shapes.

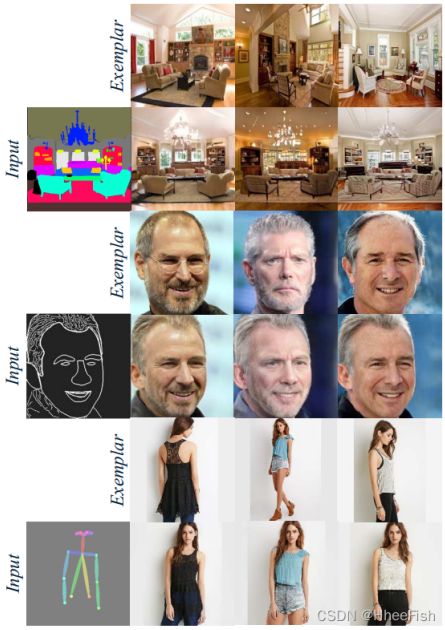


Figure 1: Result extracted from the original paper

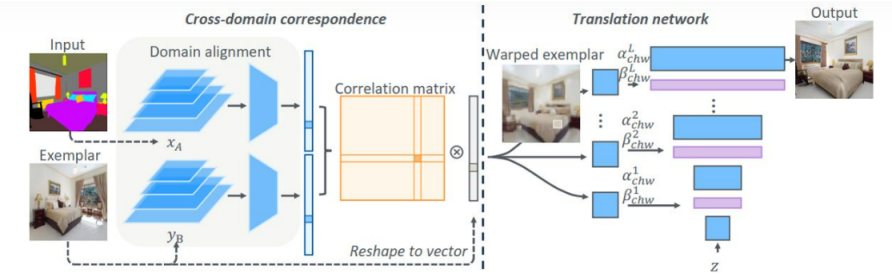


Figure 2: CocosNet architecture

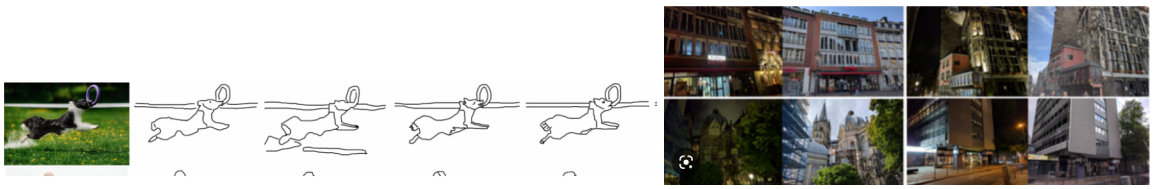
3 Approach

While the earlier models discussed have limitations, they paved the way for a more robust model named CocosNet[12]. CocosNet builds on the ideas of earlier models and combines them to enable exemplar-based image translation. In this approach, the exemplar determines the image style, while the input determines the overall structure of the target image. Unlike GAN-related models that are criticized for being out of human control, the exemplar method makes targeted image editing possible.

The network structure of CocosNet is comprised of two parts: the cross-domain correspondence network and the translation network (See Figure 2. The cross-domain correspondence network is responsible for converting images from different domains into the same domain to facilitate feature comparison. This is accomplished by utilizing a feature pyramid network to extract features of varying levels of abstraction from both the exemplar and input images, which are then transformed to a shared domain. The model then calculates the correspondence of the input and exemplar images within the shared domain to generate a warped exemplar. The translation network, on the other hand, employs techniques such as SPADE (mentioned in related work) to enhance the realism of the output image.

To constrain the output of CocosNet, the authors introduced several loss terms, such as the pseudo exemplar pairs loss, which ensures that the final output meets the desired criteria. The domain alignment loss is used to ensure that the input images are transformed into the same shared domain. The exemplar translation loss ensures that the high and low level information follows the desired requirements, while the correspondence regularization is used to promote consistency in the learned parameters. Additionally, the adversarial loss ensures that the output is realistic.

The overall framework of CocosNet is comprehensive and the results are impressive. At this point, it's difficult to see how the performance could be further improved, as the framework already combines a wealth of prior knowledge to achieve state-of-the-art results. However, one potential drawback is that the implementation is somewhat complex, and there may be opportunities to simplify the architecture and reduce the number of loss terms while still achieving similar results. As a new learner in deep learning, I still need to read more recent papers. One of my prototype ideas is to use the input image as guidance, similar to how U-net passes information of the input to the upsampling stage to guide the overall structure of the target image. However, designing appropriate loss terms to place corresponding textual information in the right position would be a challenging task. Before attempting this, I plan to explore the limitations of CocosNet by training the model on other challenging datasets. Overall, my current plan for this project is to reimplement the model and apply it to interesting applications on different datasets.



(a) Contour Drawing Dataset (b) Aachen Day-Night Dataset

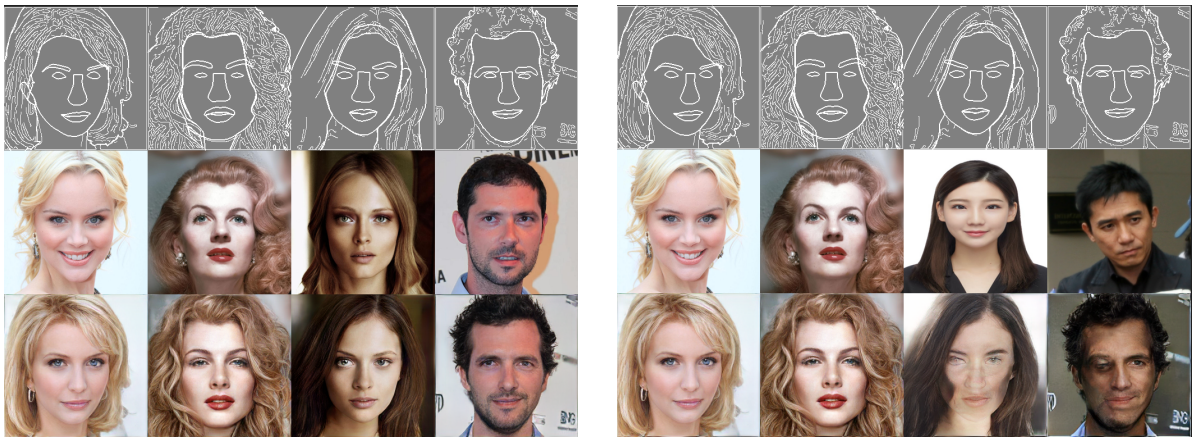
Figure 3: Datasets

4 Data

Currently, I have selected two datasets, namely the Contour Drawing Dataset and the Aachen Day-Night Dataset, for image style transfer. These datasets are relatively small, so transfer learning will be necessary to achieve optimal results. According to Pan Zhang et al, CocosNet is applicable to various tasks, and the above datasets have not been tested in the original paper. That is why I plan to explore the feasibility of using CocosNet for these tasks.

The Contour Drawing Dataset consists of over 5000 image and contour drawing pairs, with 1000 images provided, each paired with 5 drawings. The drawings are simple and contain strokes aligned with the edges of the corresponding images (as shown in Figure 1). The dataset was originally intended for training a model to convert captured photographs into contour drawings, which should be easy for CocosNet. However, I plan to attempt the more challenging task of training the model to convert contour drawings back into photographs.

The Aachen Day-Night Dataset is comprised of 14,607 photographs captured during both daytime and nighttime hours. The dataset includes numerous images of the same location taken under various weather conditions (see Figure 2). The purpose of this dataset is to train a model capable of converting a nighttime image into its corresponding daytime image.



(a) Original paper's example result

(b) My testing result

Figure 4: inference outputs

5 Preliminary Results

The CocosNet source code is relatively complex, and I am still in the process of comprehending it. However, when attempting to perform inference on new images, I noticed that the resulting output appears anomalous (as shown in Figure 4). This may be due to bias present in the training data. Additionally, it has come to my attention that the final output is also provided in the dataset. When attempting to replace the targeted output with an alternative photograph, the inference result is similarly affected. It is my belief that the target image should not be presented to the model during the inference stage. I may require additional time to determine the cause of this issue. Additionally, it appears that there are some problems with GPU53, and only GPU52 is currently accessible. The available computing resources may be somewhat limited for the project. Going forward, I intend to fully comprehend the implementation details of CocosNet and evaluate its robustness. Once this is accomplished, I will make every effort to re-implement it. Although I am uncertain if I will be able to complete the re-implementation within the given time constraints, I will endeavor to re-implement

certain key aspects of the model.

Source code of CocosNet: https://github.com/microsoft/CoCosNet

References

[1] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017.

[2] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[3] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

[7] Rui Liu, Yixiao Ge, Ching Lam Choi, Xiaogang Wang, and Hongsheng Li. Divco: Diverse conditional image synthesis via contrastive generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16377–16386, 2021.

[8] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[9] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[11] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[12] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020.

[13] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.