SiENet: Siamese Expansion Network for Image Extrapolation

Xiaofeng Zhang[a], Feng Chen[b], Cailing Wang [a], Ming Tao[a], Songsong Wu[a] and Guoping Jiang[a]

[a]*College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China*
[b]*School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China*

Abstract—Different from image inpainting, image outpainting has relatively less context in the image center to capture and more content at the image border to predict. Therefore, classical encoder-decoder pipeline of existing methods may not predict the outstretched unknown content perfectly. In this paper, a novel two-stage siamese adversarial model for image extrapolation, named Siamese Expansion Network (SiENet) is proposed. Specifically, in two stages, a novel border sensitive convolution named adaptive filling convolution is designed for allowing encoder to predict the unknown content, alleviating the burden of decoder. Besides, to introduce prior knowledge to network and reinforce the inferring ability of encoder, siamese adversarial mechanism is designed to enable our network to model the distribution of covered long range feature as that of uncovered image feature. The results on four datasets has demonstrated that our method outperforms existing state-of-the-arts and could produce realistic results. Our code is released on https://github.com/nanjingxiaobawang/SieNet-Image-extrapolation.

Index Terms—Two-stage GAN; Adaptive filling convolution; Siamese adversarial mechanism; Image outpainting.

I. INTRODUCTION

Image extrapolation, as illustrated in Fig. 1, is to generate new contents beyond the original boundaries of a given image. Even belonging to the general image painting task as inpainting [1]–[6] , image outpainting [7] has its special characteristics. We rethink this task from two aspects. First, comparing to inpainting, image extrapolation would rely on relative less context to infer much larger unknown content. Besides, the inferred content locates outside the given image. Therefore, existing expertise of inpainting can not be applied into this task directly. Second, existing methods of image extrapolation generally employ classical encoder-decoder (ED) structure. This structure forces the encoder to capture the global and local features and allows decoder to recover them to desired resolution. Namely, all the burden of inferring is put on the decoder. Yet, the input of outpainting doesn't have such abundant features, which makes the capturing ability of encoder weak and the inferring burden of decoder heavy.

Due to the requirement of outpainting task, most inpainting expertise can't be applied to it directly. [8] proposed the specific spatial expansion module and boundary reasoning module for the requirement of connecting marginal unknown content and inner context. Besides, heavy burden of prediction on decoder may result in the poor performance of generating realistic images. [9] applied DCGAN [10] which mainly

Fig. 1: Overview of difference between image inpainting and image outpainting.

focuses on the prediction ability of decoder. This kind of methods could predict the missing parts on both sides of the image, while this design would ignore coherence of semantic and content information.

In this paper, to design task-specific structure, we propose a novel two-stage siamese adversarial model for image extrapolation, named Siamese Expansion Network (SiENet). In our SiENet, a boundary-sensitive convolution, named adaptive filling convolution, is proposed to automatically infer the features of surrounding pixels outside known content with balance of smoothness and characteristic. This adaptive filling convolution is inserted to the encoder of two-stage network, activating the sensitivity of encoder for border features. Therefore, encoder could infer the unknown content and the inferring burden of decoder could be alleviated.

In addition, the siamese adversarial mechanism is designed to introduce prior knowledge into network and adjust the inferring burden of each part. In the joint training of two stages, ground truth and covered image are fed into network to calculate the siamese loss. Siamese loss encourages the features of them in the subspace to be similar, leading to reinforced predicting ability of long range encoder. Besides, a adversarial discriminator is designed in each stage to push the global generator to generate realistic prediction. Thus, the whole inferring burden of the network is reasonably allocated.

Our contributions can be summarized as follows:

- We design a novel two-stage siamese adversarial network for image extrapolation. Our SiENet, a task-specific pipeline, could regulate the inferring burden of each part and introduce prior knowledge into network legitimately.
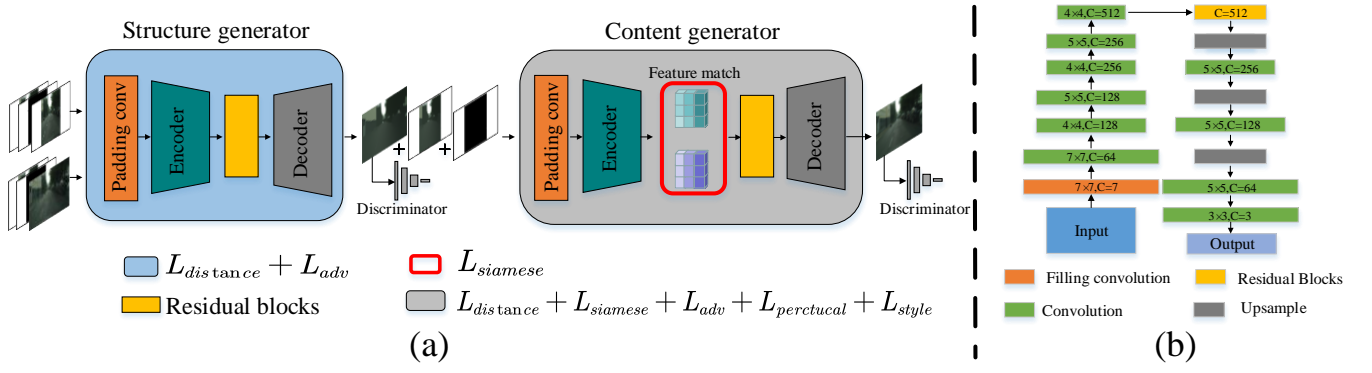
Fig. 2: Illustration of network architecture. (a) is the basic pipeline of our network. (b) is the detailed structure of structure generator, which is similar to that of content generator.

- We propose an adaptive filling convolution to concentrate on inferring pixels in unknown area. By inserting this convolution into encoder, encoder could be endowed powerful inferring ability.
- Our method achieves promising performance on four datasets and outperforms existing state-of-the-arts. The results on four datasets, i.e., Cityscapes, paris street-view, beach and Scenery, including street and nature cases, indicate the robustness of our method.

II. METHODOLOGY

Our SiENet is a two-stage network, and each stage has specific generator and discriminator, as shown in Fig. 2 (a). The details of our method are provided as follows.

A. Framework Design

Given an input image $X \in \mathbb{R}^{h \times w \times c}$, extension filling map $M \in \mathbb{R}^{h \times w \times 1}$ and smooth structure $S \in \mathbb{R}^{h \times w \times c}$ [11], our method intends to generate a visually convincing image $\widehat{Y} \in \mathbb{R}^{h \times w' \times c}$ where w' is desired width of the generated image. SieNet follows the coarse-to-fine two-stage pipeline [12]. As shown in Fig. 2 (a), an uncovered smooth structure $S_{gen} \in \mathbb{R}^{h \times w' \times c}$ is produced by the structure generator first. Then the generated structure feature S_{gen} would combine with image X and filling map M, acting as the input of second generator, i.e., content generator, to generate outstretched image $\widehat{Y}$. Besides, the discriminator D to ensure the output of generator and actual data conform to a certain distribution.

As illustrated in Fig. 2 (b), the encoder of structure generator downsamples the input with $8\times$ scale. Then to capture multi-scale information two residual blocks are designed to further proceed the output of encoder. Finally, by three nearest-neighbor interpolation, the output of residual blocks is upsampled to desired resolution. The detailed structure of content generator is similar to that of structure generator, except adding two residual blocks before last two upsampling operations. The discriminator D follows the structure and protocol of that of BicycleGAN [13], [14] .
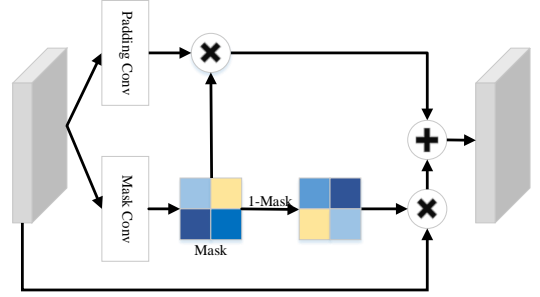

Fig. 3: Scheme of adaptive filling convolution.

B. Adaptive Filling Convolution

To endue encoder with predicting ability, we propose a boundary sensitive convolution, i.e., adaptive filling convolution. Given a convolutional kernel with K sampling locations (e.g., $K = 9$ in a 3×3 convolution), let w_k and $m_k \in [0, 1]$ denote the weight of padding convolution and mask modulation for k-th position. $x(p)$ and $y(p)$ represent the feature map at location p from input and output respectively. Therefore, the filling convolution could be formulated as follows:

$$y(p_0) = \sum_{k=1}^{K} [w_k \cdot m_k \cdot x(p_0 + p_k) + (1 - m_k) \cdot x(p_0 + p_k)] \tag{1}$$

In practical learning , the scheme of our filling convolution is shown as Fig. 3. Two separate convolutions model padding weight w_k and mask modulation m_k respectively where m_k is defaulted to 1. Our filling convolution could predict the pixel outside the image boundary. When kernel center locates just outside known content, only scant real pixels would join the patch calculation. Our padding convolution would predict it by known information with w_k, and another skip connection from input to output would keep characteristic value to avoid high variance. Besides, mask modulation m_k works as a adaptive balance factor of these two branches, to guarantee the smoothness and characteristic of prediction. We insert a 7×7 filling convolution into the bottleneck of encoder, therefore, encoder could possess the ability of capturing context and the ability of predicting unknown content.

C. Siamese Adversarial Mechanism

Classical adversarial mechanism could constrain the whole generator toward the real case. However, this constraint is implicit to each part of generator. Especially in two-stage GAN [15]–[17] for image outpainting, long range encoding may lead to insufficient inferring ability for decoder. Considering the predicting ability of encoder brought by filling convolution, we further add explicit constraint in encoder to push the features of covered and uncovered image to be common. Namely, prior knowledge of uncovered image is learned by encoder. In addition, we also take advantages of adversarial mechanism in our model to constrain the generated results to be consistent with ground truth [18], [19]. Through these two thoughtful constraints brought by our siamese adversarial mechanism, the predicting burden is well regulated.

Let G_{strut} denotes structure generator and E_{cont} denotes encoder of content generator, $I = [X, M, S]$ is the input of the whole network. The output F of long range encoder $E_{cont}(G_{strut}())$ could be formulated as:

$$F = E_{cont}(concat(X, M, G_{strut}(I))) \quad (2)$$

Therefore, let superscripts $'$ and ^{gt} denote whom is generated by covered input and ground truth input respectively, the siamese loss $L_{siamese}$ is $\ell 2$ loss of F' and F^{gt}:

$$L_{siamese} = ||F' - F^{gt}||_2 \quad (3)$$

In order to generate realistic results, we introduce adversarial loss L_{adv} in the structure generator:

$$L_{adv} = \mathbb{E}[\log(1 - D(\widehat{Y})] + \mathbb{E}[\log D(Y)] \quad (4)$$

D. Loss Function

We introduce $L_{distance}$ loss to predict the distance between the structures S'_{gen} and S^{gt}:

$$L_{distance} = \left\| S'_{gen} - S^{gt} \right\|_1 \quad (5)$$

Besides, the perceptual loss $L_{perctual}$ and the style loss L_{style} [20] are applied into our network. $L_{perctual}$ is defined as:

$$L_{perctual} = \mathbb{E}\left[\sum_i \frac{1}{N_i} \left\| \phi_i(Y) - \phi_i\left(\widehat{Y}\right) \right\|_1 \right], \quad (6)$$

ϕ_i is the activation map of the i-th layer of the pre-trained network. In our work, ϕ_i is activation map from the layers of relu1-1, relu2-1, relu3-1, relu4-1 and relu5-1 of the ImageNet pre-trained VGG-19. These activation maps are also used to calculate the style loss to measure the covariance between the activation maps. Given a feature map size $C_j * H_j * W_j$, the style loss L_{style} is calculated as follows.

$$L_{style} = \mathbb{E}\left[\left\| G^\phi\left(\widehat{Y}\right) - G^\phi(Y) \right\|_1 \right], \quad (7)$$

G^ϕ is the the reconstructed Gram metric based on activation map ϕ_j. Totally, the overall loss L_{total} is:

$$L_{total} = \lambda_{dist}L_{distance} + \lambda_{adv}L_{adv} + \lambda_p L_{perctual} + \lambda_s L_{style} + \lambda_{sie}L_{siamese}, \quad (8)$$

where λ_{dist}, λ_{adv}, λ_p, λ_s and λ_{sie} are set 5, 1, 0.1, 250 and 1 respectively.

III. EXPERIMENTS

A. Implementation Details and Configurations

Our method is flexible for two-direction and single-direction outpainting. For two-direction outpainting, our method is evaluated on three datasets, i.e., Cityscapes [21], paris street-view [2] and beach [9]. Single-direction evaluation is made on Scenery dataset [22]. Besides, we take structural similarity (SSIM), peak signal-to-noise ratio (PSNR) as our evaluation metrics.

In training and inference, the images are resized to 256×256. We train our model using Adam optimizer [23] with learning rate 0.0001, beta1 0 and beta2 0.999. The batch size is set to 8 for paris-street view and beach, 2 for Cityscapes, and 16 for Scenery. The total iteration of joint training is 10^6 for four datasets.

TABLE I: Ablation experiments of filling convolution and siamese adversarial mechanism on Cityscapes. 'FConv' and 'SAM' denote filling convolution and siamese adversarial mechanism respectively.

FConv	SAM	SSIM	PSNR
		0.6896	23.2539
√		0.7254	23.5987
	√	0.7647	24.4578
√	√	0.7832	24.8213

B. Ablation Study on SiENet

C. Comparisons

To validate the effectiveness of filling convolution and siamese adversarial mechanism in our network, We make the ablation experiment in Tab. I. Only using filling convolution or siamese adversarial mechanism could get a significant improvement over classical encoder-decoder. And the combination of them could achieve the best performance, indicating the regulation of predicting burden is beneficial for the task.

Fig. 4: Ablation study of effectiveness of filling convolution and siamese adversarial mechanism on Cityscapes and beach.

TABLE II: Performance comparison of image two-direction and single-direction outpainting. The two-direction outpainting is evaluated on beach, paris-street view and Cityscapes, and single-direction outpainting is evaluated on Scenery.

Datasets / Methods	SSIM				PSNR			
	Beach	Paris	Cityscapes	Scenery	Beach	Paris	Cityscapes	Scenery
Image-Outpainting	0.3385	0.6312	0.7135	-	14.6256	19.7400	23.0321	-
Outpainting-srn	0.5137	0.6401	0.7764	-	18.2211	19.2734	23.5927	-
Edge-Connect	0.6373	0.6342	0.7454	-	19.8372	22.1736	24.1413	-
NS-Outpainting	-	-	-	0.6763	-	-	-	17.0267
Our method	**0.6463**	**0.6428**	**0.7832**	**0.8557**	**20.7965**	**23.9794**	**24.8213**	**31.7686**

Fig. 5: Quantitative two-direction comparisons of our method with our state-of-the-arts on Cityscapes.
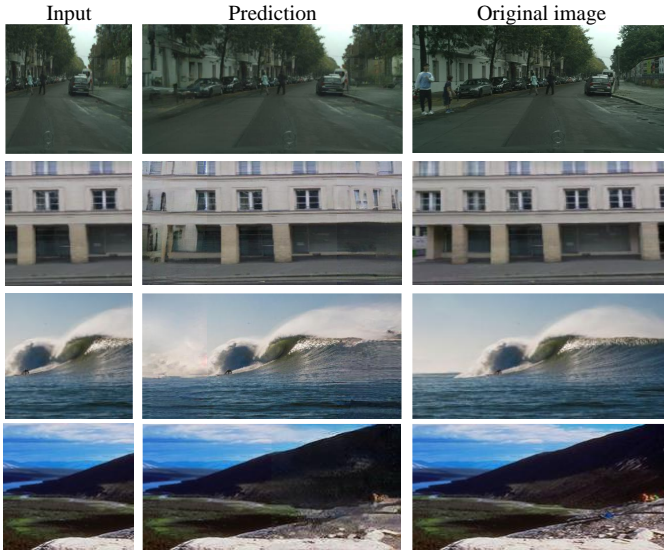

Fig. 6: The visual performance of our method on four dataset. The front three rows represent two-direction outpainting results. The last row is single-direction outpainting result.

We also visualize the results of street and nature scenery to measure the quantitative capability of them. As shown in Fig. 4, the combination of these proposed two parts could recover the details and generate realistic results in easy nature and complicated street cases.

Our method is compared with existing methods including Image-Outpainting [9], Outpainting-srn [8], Edge-connect [24] and NS-outpainting [22]. As shown in Tab. II, for two-direction outpainting, our method outperforms Image-Outpainting [9], Outpainting-srn [8], Edge-connect [24], and achieves state-of-the-art performance in three datasets. For single-direction, the great margin of performance between our method and NS-outpainting [22] indicates the superiority and generality of our SiENet. Besides, we also provide visual comparison of various methods in Fig. 5 and visual performance of our method on four datasets in Fig 6. Notably, the images generated by our method are similar to the ground truth than that of existing methods.

IV. CONCLUSION

A novel end-to-end model named Siamese Expansion Network for Image Extrapolation is propoesd in this paper. To regulate the heavy predicting burden of decoder legitimately, we first propose adaptive filling convolution to endow encoder with predicting ability. Then we introduce siamese adversarial mechanism in long range encoder to allow encoder to learn prior knowledge of real image and reinforce the inferring ability of encoder. Our method has achieved promising performance on four datasets and outperforms existing state-of-the-arts.

REFERENCES

[1] Barnes C, Shechtman E, Finkelstein A, et al. PatchMatch: A randomized correspondence algorithm for structural image editing[J]. ACM Transactions on Graphics, 2009, 28(3): 24.

[2] Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: Feature learning by inpainting[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016: 2536-2544.

[3] Yu J, Lin Z, Yang J, et al. Generative image inpainting with contextual attention[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 5505-5514.

[4] Yang C, Lu X, Lin Z, et al. High-resolution image inpainting using multi-scale neural patch synthesis[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 6721-6729.

[5] Cho J H, Song W, Choi H, et al. Hole filling method for depth image based rendering based on boundary decision[J]. IEEE Signal Processing Letters, 2017, 24(3): 329-333.

[6] Sulam J, Elad M. Large inpainting of face images with trainlets[J]. IEEE Signal Processing Letters, 2016, 23(12): 1839-1843.

[7] Ledig C, Theis L, Huszr F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 4681-4690.

[8] Wang Y, Tao X, Shen X, et al. Wide-context semantic image extrapolation[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2019: 1399-1408.

[9] Sabini M, Rusak G. Painting outside the box: Image outpainting with gans[J]. arXiv preprint arXiv:1808.08483, 2018.

[10] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.

[11] Xu L, Yan Q, Xia Y, et al. Structure extraction from texture via relative total variation[J]. ACM Transactions on Graphics, 2012, 31(6): 1-10.

[12] Ren Y, Yu X, Zhang R, et al. StructureFlow: Image inpainting via structure-aware appearance flow[C]//Proceedings of the IEEE International conference on Computer Vision. 2019: 181-190.

[13] Zhu J Y, Zhang R, Pathak D, et al. Toward multimodal image-to-image translation[C]//Advances in Neural Information Processing Systems. 2017: 465-476.

[14] Iizuka S, Simo-Serra E, Ishikawa H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics, 2017, 36(4): 1-14.

[15] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in Neural Information Processing Systems. 2014: 2672-2680.

[16] Arjovsky, Martin, Soumith Chintala, and Lon Bottou. "Wasserstein gan." arXiv preprint arXiv:1701.07875, 2017.

[17] Gulrajani, Ishaan, et al. "Improved training of wasserstein gans." Advances in Neural Information Processing Systems. 2017: 5767-5777.

[18] Norouzi M, Fleet D J, Salakhutdinov R R. Hamming distance metric learning[C]//Advances in Neural Information Processing Systems. 2012: 1061-1069.

[19] Melekhov I, Kannala J, Rahtu E. Siamese network features for image matching[C]//2016 23rd International Conference on Pattern Recognition. 2016: 378-383.

[20] Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016: 2414-2423.

[21] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016: 3213-3223.

[22] Yang Z, Dong J, Liu P, et al. Very Long Natural Scenery Image Prediction by Outpainting[C]//Proceedings of the IEEE International Conference on Computer Cision. 2019: 10561-10570.

[23] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[24] Nazeri K, Ng E, Joseph T, et al. Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv 2019[J]. arXiv preprint arXiv:1901.00212.