Structure-Aware Image Expansion with Global Attention

Dewen Guo
Peking University
guodewen@pku.edu.cn

Jie Feng
Peking University
feng_jie@pku.edu.cn

Bingfeng Zhou
Peking University
cczbf@pku.edu.cn

ABSTRACT

We present a novel structure-aware strategy for image expansion which aims to complete an image from a small patch. Different from image inpainting, the majority of the pixels are absent here. Hence, there are higher requirements for global structure-aware prediction to produce visually plausible results. Thus, treating the expansion tasks as inpainting from the outside is ill-posed. Therefore, we propose a learning-based method combining structure-aware and visual attention strategies to make better prediction. Our architecture consists of two stages. Since visual attention cannot be taken full advantage of when the global structure is absent, we first use the ImageNet-pre-trained VGG-19 to make the structure-aware prediction on the pre-training stage. Then, we implement a non-local attention layer on the coarsely-completed results on the refining stage. Our network can well predict the global structures and semantic details from small input image patches, and generate full images with structural consistency. We apply our method on a human face dataset, which containing rich semantic and structural details. The results show its stability and effectiveness.

CCS CONCEPTS

• **Computing methodologies** → *Computational photography*; *Image processing*.

KEYWORDS

Image expansion, structure-aware, global attention, generative adversarial network

ACM Reference Format:
Dewen Guo, Jie Feng, and Bingfeng Zhou. 2019. Structure-Aware Image Expansion with Global Attention. In *SIGGRAPH Asia 2019 Technical Briefs (SA '19 Technical Briefs), November 17–20, 2019, Brisbane, QLD, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3355088.3365161

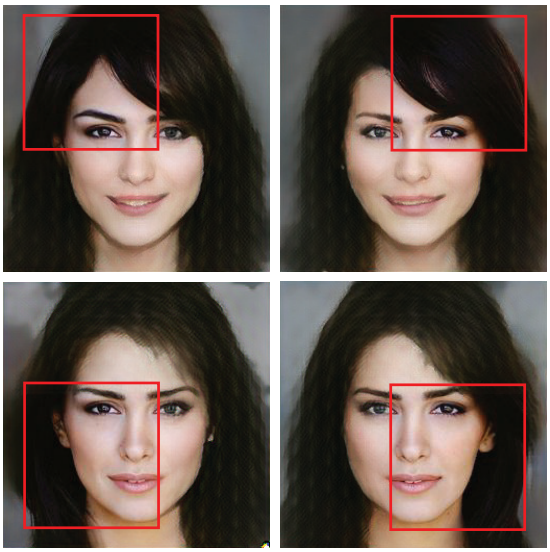

Figure 1: Image expansion. The outputs are generated from small patches extracted from different spatial locations of the same original image. Our method can produce the expansions with reasonable structure. The spatial location of the reference patches are indicated with red boxes in the output images.

1 INTRODUCTION

Image expansion can be thought as complete an image from the outside while maintaining the semantic and structural coherency. Traditional image expansion methods provide conceptually simple thoughts of real image data manipulation such as database-driven extrapolation[Wang et al. 2014] and panorama stitching[Brown and Lowe 2007]. Recently, learning based algorithms such as image outpainting[Sabini and Rusak 2018], Semantic Regeneration Network (SRN)[Wang et al. 2019] and adversarial texture expansion[Zhou et al. 2018] introduce the Generative Adversarial Networks (GANs) to such tasks.

In recent research works, various classic image inpainting methods are applied in image expansion. Contextual attention method[Yu et al. 2018] opened up new frontiers in image inpainting utilizing spatially distant contextual information. With such visual attention mechanism, local convolutional operators are able to percept similar features extracted from distant spatial locations. Afterwards, several kinds of attention masks are introduced to obtain better results.

It is relatively simple to generate coarse results with structural coherency in inpainting tasks, since the small absent regions are usually inside the middle of the images, with rich contextual and structural information around them. For instance, vanilla GAN with attention and local-global consistency[Iizuka et al. 2017] may produce nice results.

To solve the problem of the structure information scarcity in image expansion, we leverage perceptual features[Gatys et al. 2016; Johnson et al. 2016] when constructing our regularization to predict coarse results with strong structure-aware features. Therefore,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SA '19 Technical Briefs, November 17–20, 2019, Brisbane, QLD, Australia
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6945-9/19/11…$15.00
https://doi.org/10.1145/3355088.3365161

we are able to use the features borrowed by the global-attention layer from the synthesized coarse results in spatially distant regions. To stabilize the training procedure, our network architecture utilize some recent training strategies and module designs, such as coarse-to-fine architecture, Wasserstein GAN with gradient penalty (WGAN-GP)[Gulrajani et al. 2017] and Relative Spatial Variant (RSV) masks[Wang et al. 2019].

Our contributions are summarized as follows.

- We present an end-to-end GAN architecture for image expansion. To our knowledge, it is the first network that introduce the attention mechanism to image expansion tasks.
- We provide a structure-aware regularization to maintain the quality of the output results. The regularization term acts as a dominant building block in our method.

2 PROPOSED METHOD

Our goal is to rebuild a structure-plausible image based only on a small patch of the original image.

Due to the absence of most pixels, visual attention mechanism cannot be directly implemented on the expansion tasks. To address this, we firstly predict the structure of each image patch, then introduce a visual attention module to enhance the output quality.

As overviewed in Fig. 2, our network architecture adopts a 2-stage training strategy. The first stage, i.e. the pre-training stage, aims to generate a structure-aware guidance for the following refining stage. The first stage is an encoder-decoder convolutional architecture with skip connections between the counterparts with identical scales in both ends. Our motivation of a two-stage training strategy is to let the architecture predict the possible global structure from a relatively small given patch. Different from recent state-of-the-art[Wang et al. 2019], we directly use VGG features to regularize the structural prediction instead of a Markov random field (MRF). On the second stage, a refining network is appended to the pre-training module and both are trained jointly to produce final results. We introduce a global attention layer to the refining stage inspired by non-local nets and visual self-attention[Wang et al. 2018; Zhang et al. 2018].

2.1 Structure-Aware Regularization

Usually, implementing pixel-wise loss on RGB images lacks consideration of the global structure. To assess the perceptual differences between the synthesized results and the original images, we utilized the feature maps extracted by pre-trained VGG-19 in our regularization term.

Different layers of VGG-19 focus on different kinds of details and patterns. Initial convolutional layers of VGG-19 are able to reconstruct the images perfectly. However, the reconstruction quality decays as the processing flow going deeper in the network. In deeper layers of the net, dilated pixel details are neglected while the general structure information are preserved[Gatys et al. 2016]. Similarly, style features can also be extracted from the net. We construct our regularization from different sublayers of the net. To balance the effect among the structure-aware regularization, the adversarial training and the detail regression, different coefficients of the regularization term are set in different stages of our training procedure. Based on empirical knowledges, we calculate $\mathcal{L}_1$ rather than Mean Square Error (MSE) differences between the source and target feature maps to prevent the reconstructions from yielding blurry results. The structure-aware regularization term is formulated as in Eq. 1,

$$\begin{aligned}\mathcal{L}_{GS} =&\lambda_{cs} \left\|\mathcal{V}_{cs}(f(x)) - \mathcal{V}_{cs}(O)\right\|_1 + \\ &\lambda_s \left\|\mathcal{V}_s(f(x)) - \mathcal{V}_s(O)\right\|_1,\end{aligned} \tag{1}$$

where $\mathcal{V}_{cs}$ is the content- / structure-representation layer of VGG-19, and $\mathcal{V}_s$ is the style-representation layer.

2.2 Global Attention Modeling

In the refining model, dilated convolution is adopted to expand the receptive field, because the standard convolution is a local operation whose receptive field depends only on the kernel size. Visual attention mechanisms construct the dependencies among spatially distant yet relevant pixels. Recent researches[Yu et al. 2018] introduce this mechanism to inpainting tasks where they are only relatively small-sized absent regions. Here, we introduce a *global attention* layer to accomplish image expansion tasks, even though most pixels are absent. Inspired by non-local nets and visual self-attention, our global-attention map can be formulated as:

$$\mathcal{M}_A = f(x) \otimes \mathcal{S}(x^T w_\theta^T w_\phi x), \tag{2}$$

where $f(\cdot), \theta$ and ϕ indicate 1×1 convolution. The calculation is demonstrated in Fig. 3. Here $\mathcal{S}$ is the softmax operation, and $\otimes$ indicates the matrix multiplication. To be specific, we utilize the embedded Gaussian function[Wang et al. 2018] $\mathcal{EG}(\cdot,\cdot)$ for the softmax computation (Eq. 3).

$$\mathcal{EG}(x_i, x_j) = \exp(\theta(x_i)^T \phi(x_j)). \tag{3}$$

Hence, the global attention is formed as

$$\mathcal{S}(\theta(x_i)^T \phi(x_j)) = \frac{\mathcal{EG}(x_i, x_j)}{\Sigma_j \mathcal{EG}(x_i, x_j)}. \tag{4}$$

The global attention mechanism aims to utilize the feature patches spatially distant from the local convolution operations. More general structure details could be learnt by such an attention layer. After calculating the attention map, the contribution score of each pixel to the current local convolution will guide the synthesis of the image.

2.3 Learning Objectives

We adopt the WGAN-GP[Gulrajani et al. 2017] as our basic architecture. The adversarial loss can be demonstrated as:

$$\begin{aligned}\mathcal{L}_{adv} =& -\lambda_D \mathbb{E}_{x\sim\mathbb{P}_x}[\log D(G(x))] + \\ &\lambda_\nabla \mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x}) \odot M\|_2 - 1)^2].\end{aligned} \tag{5}$$

Here M is the mask to indicate the locations of the lost pixels. The latter term of the loss function is the gradient penalty that penalizing the $\|\nabla_{\hat{x}} D(\hat{x}) \odot M\|_2$ if it is near to 1 to stabilize the model. Intuitively speaking, we want the distribution of $G(\hat{x})$ as close as possible to x, while $D(G(\hat{x}))$ cannot overpass $D(x)$.

Considering both local and global consistency, and structure-aware regularization, the final objective is formulated as:

$$\begin{aligned}\mathcal{L} =&\lambda_L \left\|M \odot (\hat{x} - x)\right\|_1 + \lambda_G \left\|\hat{x} - x\right\|_1 + \\ &\lambda_{adv}\mathcal{L}_{adv} + \lambda_{GS}\mathcal{L}_{GS}.\end{aligned} \tag{6}$$

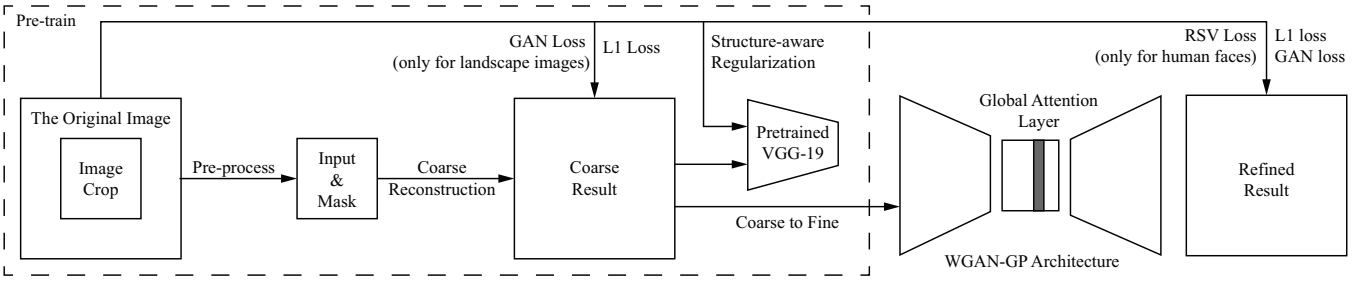

Figure 2: Our network architecture. The training procedure is divided into two phrases, namely pre-training and refinement training (full training). The objective functions are depicted in the figure. We use WGAN-GP for stabilizing the training process. In the middle of the refinement network is a global-attention layer guiding the contribution of each pixel in synthesis of spatially different locations.

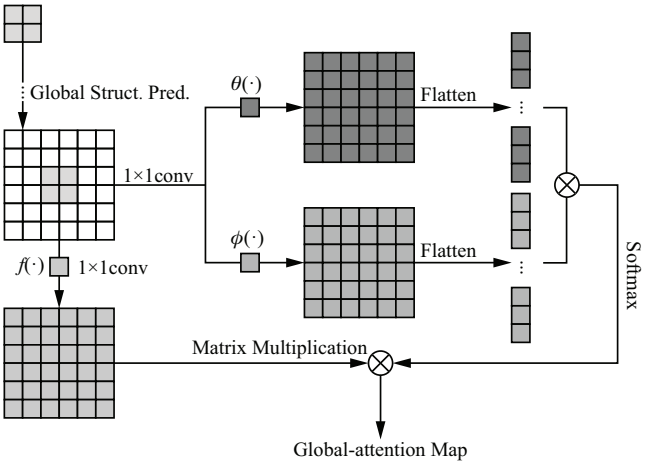

Figure 3: Global-attention map generation.

3 EXPERIMENTS

3.1 Dataset and System Configuration

We implement our method on CelebA-HQ[Karras et al. 2017] dataset. For visual evaluation, we retrain the SRN[Wang et al. 2019] model by running the open source codes on the same dataset. Due to the limitation of GPU RAM, the training batch size is set to be 8 for both models.

Moreover, we also apply our method on some landscape image datasets, including landscape images dataset collected from Places2[Zhou et al. 2017] and CycleGAN[Zhu et al. 2017].

Mentioned models are trained on an NVIDIA Titan X GPU with 12GB of RAM.

3.2 Training Procedure

3.2.1 Pre-training without attention. The main challenge in the image expansion tasks is the lacking of the most of the structure information. When only a minority of the pixels are missing, we can make predictions for them based on empirical knowledges that roughly infer the position and structure of missing parts. But, when given only a small fraction of the image, making prediction of the whole image would be much more difficult. A naïve solution is to consider the image expansion task as inpainting outside the boundaries. However, that would cause structural artifacts on the second training stage.

Our pre-training stage aims to make the network learn the structural-level prediction. Reasonable structure prediction works as a global guidance in the second training stage.

For this purpose, we set the structure-aware regularization term and increase its weight in our pre-training objective function. The pre-training results of our model contain more structural clues in comparison with those of SRN, which are demonstrated in Fig. 4.

3.2.2 Full training with attention layer. In our architecture, the global attention layer is a redesigned version of the non-local block and the self-attention layer, based on the coarse image reconstructions from the corresponding patches. The full training is similar to the coarse-to-fine architectures while introducing the attention mechanism into the image expansion tasks. The comparisons between the results from the retrained SRN and ours are demonstrated in Fig. 4. Notice the parts of *forehead*, *nose*, *jawline* and *mouth* in the images respectively, our method can produce structurally sound results of different facial parts. More comparisons are shown in the supplemental material.

3.2.3 Application in natural scene images expansion. The natural scene images are more complex, and its expansion is even more challenging because the distributions of the pixel intensity can be so various among different images.

To solve this problem in the case of natural scenes, we fine-tune our network by adding the generative loss term while giving up the structure-aware regularization on the refining stage. The images are expanded horizontally, which are shown in Fig. 5. Our method can expand the natural scene images with either structural coherence or realistic texture details.

4 CONCLUSIONS AND FUTURE WORK

In this work, we propose a systematic structure-aware image expansion framework with global attention. We explore the potential global structure information to reconstruct better results in image expansion tasks. The global attention is beneficial in both structure prediction and receptive field expanding. Combining the structure-aware regularization with global attention, our method achieves

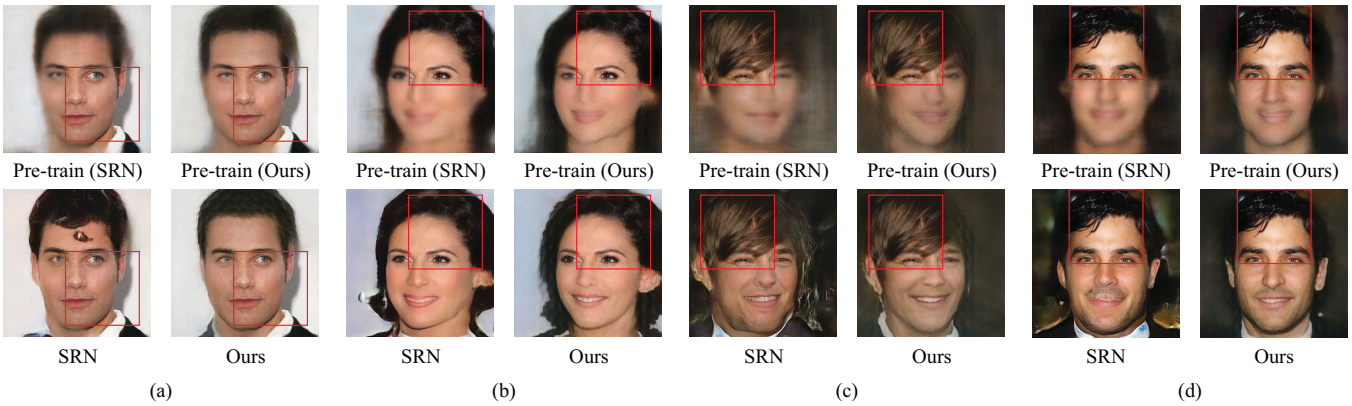

Figure 4: The comparisons between retrained SRN[Wang et al. 2019] and our method. Both networks are trained on the same face dataset. Our pre-training results show more structural details and thus output more structurally sound predictions in final expanded images. The spatial location of the reference patches are indicated with red boxes in the output images.

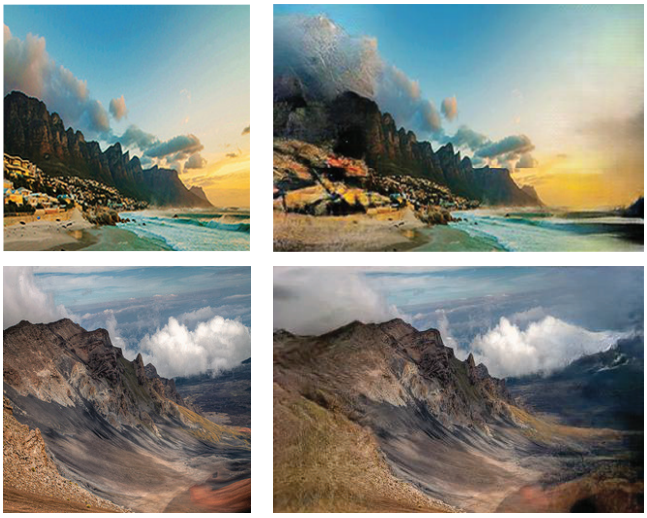

Figure 5: Natural scene images expansion. Fine-tuning our network can tackle different kinds of expansion tasks. (Input images courtesy of CycleGAN[Zhu et al. 2017].)

structurally sound results. In the future, we may expand the images on various kinds of challenging data such as natural scenes, different animals or other objects. The synthesis quality of the high frequency details such as human hair should also be improved in the future work. Furthermore, to get photorealistic results with plausible boundary details needs higher level of feature-perception mechanism, which is a prospective field of research.

ACKNOWLEDGMENTS

We appreciate the anonymous reviewers for their suggustions. This work was supported by National Natural Science Foundation of China (NSFC) [grant number 61872014], National Key Research and Development Program of China [grant number 2016QY02D0304] and Seengene Inc. [Contract No.2019110016000167].

REFERENCES

Matthew Brown and David G Lowe. 2007. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision (IJCV)* 74, 1 (2007), 59–73.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2414–2423.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NIPS)*. 5767–5777.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and Locally Consistent Image Completion. *ACM Trans. Graph. (TOG)* 36, 4, Article 107 (July 2017), 14 pages. https://doi.org/10.1145/3072959.3073659

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*. Springer, 694–711.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv preprint arXiv:1710.10196* (2017).

Mark Sabini and Gili Rusak. 2018. Painting Outside the Box: Image Outpainting with GANs. *CoRR* abs/1808.08483 (2018).

Miao Wang, Yu-Kun Lai, Yuan Liang, Ralph R. Martin, and Shi-Min Hu. 2014. BiggerPicture: Data-driven Image Extrapolation Using Graph Matching. *ACM Trans. Graph. (TOG)* 33, 6, Article 173 (Nov. 2014), 13 pages. https://doi.org/10.1145/2661229.2661278

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7794–7803.

Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. 2019. Wide-Context Semantic Image Extrapolation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2018. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5505–5514.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2018. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318* (2018).

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2017).

Yang Zhou, Zhen Zhu, Xiang Bai, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2018. Non-stationary Texture Synthesis by Adversarial Expansion. *ACM Trans. Graph. (TOG)* 37, 4, Article 49 (July 2018), 13 pages. https://doi.org/10.1145/3197517.3201285

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networkss. In *IEEE International Conference on Computer Vision (ICCV)*.