SCAN: Spatial and Channel Attention Normalization for Image Inpainting

Shiyu Chen[1], Wenxin Yu[1(✉)], Liang Nie[1], Xuewen Zhang[1], Siyuan Li[1], Zhiqiang Zhang[1], and Jun Gong[2]

[1] Southwest University of Science and Technology, Mianyang, China
yuwenxin@swust.edu.cn
[2] Beijing Institute of Technology, Beijing, China

Abstract. Image inpainting focuses on predicting contents with shape structure and consistent details in damaged regions. Recent approaches based on convolutional neural network (CNN) have shown promising results via adversarial learning, attention mechanism, and various loss functions. This paper introduces a novel module named Spatial and Channel Attention Normalization (SCAN), combining attention mechanisms in spatial and channel dimension and normalization to handle complex information of known regions while avoiding its misuse. Experiments on the varies datasets indicate that the performance of the proposed method outperforms the current state-of-the-art (SOTA) inpainting approaches.

Keywords: Image inpainting · Deep learning · Attention mechanism · Normalization

1 Introduction

Image inpainting task as a research hotspot in computer vision has many applications in photo editing, object removal, and image-based rendering. The attention mechanism is becoming increasingly popular as a plug-and-play module to encode where to emphasize or suppress. The attention mechanism [5,9] in spatial dimension explicitly constructs the long-range dependency between the pixels or regions inside and outside the hole via computing their similarity to tackle the ineffectiveness of basic convolutional neural networks in learning long-range information. Also, some approaches model the attention map in channel dimension to enhance those critical features. Feature normalization (FN) is an important technique to help neural network training, typically normalizing features across spatial-dimension, even though it can lead to the shift of mean and variance and mislead training due to the impact of information in damaged regions.

Motivated by these, we proposed a two-stage network with Spatial and Channel Attention Normalization (SCAN) module to handle these problems. The SCAN module we proposed comprises Spatial Attention Normalization (SAN) and Channel Attention Normalization (CAN). The idea of SAN, same as region normalization [11] and attentive normalization [6], is to improve instance normalization. The idea of CAN is to improve group normalization [8].

T. Mantoro et al. (Eds.): ICONIP 2021, CCIS 1517, pp. 674–682, 2021.
https://doi.org/10.1007/978-3-030-92310-5_78

The models we presented are evaluated on the test dataset of CelebA-HQ and Paris Street View. Compared with those SOTA, the produced results achieve significant improvement. Our main contributions are as follows:

- We propose a Spatial Attention Normalization method for image inpainting, which will not be disturbed by damaged information when normalization.
- We propose a Channel Attention Normalization method that can strengthen the connection between similar semantics and enable semantic separation.
- Experiments on the CelebA-HQ and Paris Street View datasets show the superiority of our approach compares to the existing advanced approaches.

2 Related Work

2.1 Learning-Based Image Inpainting

The methods based on convolutional neural networks [3,4] are introduced to help understand the semantic of images during inpainting in the last few years. Pathak et al. [4] first introduce Context Encoder, which uses an encoder-decoder architecture with adversarial training to analyse the high-level semantic information. Nazeri et al. [3] divide the training process into two parts, taking the outputs of the first stage as prior structure knowledge to guide image inpainting in the second stage.

However, these approaches let convolution kernels deal with the information inside and outside the hole regions in the same way, which will mislead the encoder. Liu et al. [2] take this issue via exploiting the partial convolutional layer and mask-update operation. After then, Yu et al. [10] present the Gated Convolutional that learns a dynamic mask-update mechanism to replace the hard mask-update.

2.2 Attention Mechanism

Yu et al. [9] propose a contextual attention layer to catch and borrow information of the related patches explicitly at distant spatial locations from the hole. Wang et al. [5] propose a multi-scale attention module to capture information in multiple scales via using attention module of different matching patch sizes.

Meanwhile, the interdependencies between channels of feature map in deep layers are also important for the model to understand the semantics of the image. Hu et al. [1] introduce a Squeeze-and-Excitation module to calculate the relationship between channels by exploiting the averages of each channel explicitly. Woo et al. [7] aggregate channel information of a feature map by using max and average pooling operations and then forward them to a shared MLP.

2.3 Normalization

In the inpainting task, Tao et al. [11] exploit Region Normalization to simply separates the feature map into only two regions, uncorrupted region, and

corrupted region, according to region mask and normalize them separately to avoid the effect of error information in holes. Yi et al. [6] further propose Attentive Normalization (AN) to divide the feature map into different semantic regions by the learning semantic map and normalize them separately. AN cannot avoid the possibility of being misled by error information in damaged regions and invalid information in known regions because AN randomly selects n feature pixels from translated feature map as initial semantic filters to guide the learning of the semantic layout map.

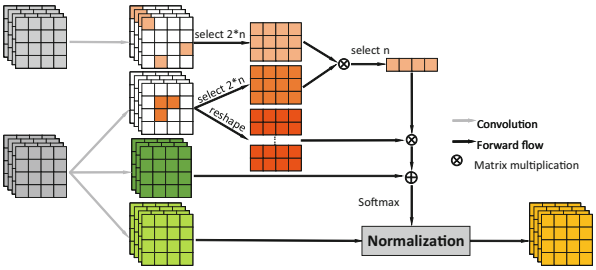

Fig. 1. The proposed SAN module. The selected patches' size is 3×3.

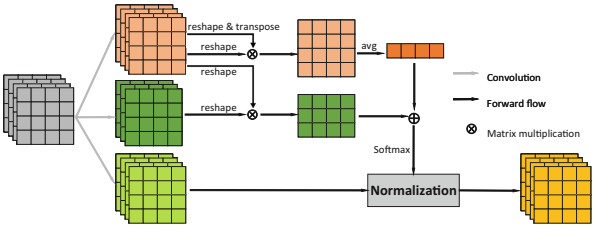

Fig. 2. The proposed CAN module.

3 Method

3.1 Spatial Attention Normalization (SAN)

The SAN can be divided into three steps, attention map learning, self-sampling regularization, and normalization, as shown in Fig. 1.

Attention Map Learning. For the given feature map $\mathbf{X} \in \mathbb{R}^{c \times h \times w}$ from decoder, we have filters $K_s \in \mathbb{R}^{n \times c \times 1 \times 1}$ to fit the semantic entities, where n denotes a predefined number of semantics entities.

We define the correlation between the feature map and these semantic entities as the raw attention map about these semantic entities $\mathbf{S}^{raw}$, of which the calculation could be implemented as a convolution calculation $\mathbf{S}^{raw} = K_s(\mathbf{X})$.

Further, to ensure that these filters learns diverse semantic entities, orthogonal regularization is employed to these entities as

$$\mathcal{L}_{so} = \lambda_{so} \| K_s K_s^T - \mathrm{I} \|_{\mathrm{F}}^2 \tag{1}$$

where $K_s \in \mathbb{R}^{n \times c}$ is a weight matrix squeezed to 2-dimension.

Self-sampling Regularization. We firstly randomly select 2*n patches from damaged regions of feature map $\mathbf{X}$ and known regions of $\mathbf{F}$, respectively, where $\mathbf{F}$ is the feature maps from the encoder located symmetrically to $\mathbf{X}$. We use partial convolution [2] to update the mask to distinguish the damaged regions from known regions and denote the selected patches from them as $s(\mathbf{X})^{2n}$ and $s(\mathbf{F})^{2n}$, respectively. Then the correlation between them could be calculated as

$$\varphi_{i,j} = < \frac{s(\mathbf{X})_i^{2n}}{\| s(\mathbf{X})_i^{2n} \|}, \frac{s(\mathbf{F})_j^{2n}}{\| s(\mathbf{F})_j^{2n} \|} > \tag{2}$$

where i and j denote index of patches.

Next, we select the n patches in $s(\mathbf{F})^{2n}$ with the highest similarity to the damaged regions as regularizing semantic filters $s(\mathbf{F})^n$, according to the relation-map φ, distinguishing helpful semantic information. Finally, the calculation of regularization term $\mathbf{S}^{re}$ could be implemented using $s(\mathbf{F})^n$ as kernels to perform convolution calculations on $\mathbf{X}$.

Normalization. Then, the semantics attention map $\mathbf{S}$ are computed as $\mathbf{S} = \mathbf{S}^{raw} + \alpha \mathbf{S}^{re}$, where $\alpha \in \mathbb{R}^{1 \times 1 \times n}$ is a learnable vector initialized as 0.1. It adjusts the effects of $\mathbf{S}^{re}$ adaptively, preventing some entities from learning useless semantics. In order to get attention score map $\mathbf{S}^*$, we apply the softmax operations in channel dimension. Each $\mathbf{S}_i^*$ (i is the index of channels) is a soft weight map, indicating the probability of every pixel belonging to semantic entity i.

According to the attention score map, we can divide the feature map into n regions and normalization them respectively as

$$\bar{\mathbf{X}} = \sum_{i=1}^{n} \frac{\mathbf{X} - \mu(\mathbf{X}_{\mathbf{S}_i^*})}{\sigma(\mathbf{X}_{\mathbf{S}_i^*}) + \epsilon} \odot \mathbf{S}_i^* \tag{3}$$

where $\mathbf{X}_{\mathbf{S}_i^*} = \mathbf{X} \odot \mathbf{S}_i^*$ and broadcast operations first broadcast $\mathbf{X}$ and $\mathbf{S}^*$ to $\mathbb{R}^{c \times n \times h \times w}$ to match the dimensions of the two matrices. $\mu(\cdot)$ and $\sigma(\cdot)$ compute the mean and standard deviation from instance respectively. Each region shares a mean and variance, strengthening the connection between internal pixels, even if they are far apart.

The final output of SAN computes the weighted sum of the original input feature map and the normalized one as $SAN(\mathbf{X}) = \gamma \bar{\mathbf{X}} + \mathbf{X}$, where γ is a learnable scalar initialized as 0.

3.2 Channel Attention Normalization (CAN)

The commonly used group normalization (GN) [8] only simply groups the channels. However, CAN groups these channels adaptively through semantic similarity and operate normalization within the group. Similar to SAN, CAN can be divided into three same steps, as shown in Fig. 2.

Attention Map Learning. Firstly, we calculate the attention map $\mathbf{C}$ by convolution calculation using K_c as filters, similar to $\mathbf{S}^{raw}$. We still have $\mathcal{L}_{co}$ to guide K_c, which is calculated similarly to Eq. (1). In order to divide the channels of the feature map into various groups according to their semantics, we continue to calculate the correlation between $\mathbf{X}$ and the attention map $\mathbf{C}$ as the raw grouping basis $\mathbf{R}^{raw}$. Specifically, we reshape $\mathbf{C}$ to $\mathbb{R}^{n\times hw}$ and $\mathbf{X}$ to $\mathbb{R}^{c\times hw}$, and then perform a matrix multiplication between $\mathbf{C}$ and the transpose of $\mathbf{X}$.

Self-attention Sampling. Different from SAN, we first calculate the correlation between channels and then calculate its average value as

$$\mathbf{R}_j^{re} = \frac{\sum_{i=0,i\neq j}^{c} \mathbf{X}_i(\mathbf{X}_j)^T}{c} \tag{4}$$

$\mathbf{R}_j^{re}$ represents the average correlation between the j-th channels and other channels, $j \in [1, c]$. $\mathbf{R}^{re}$ is broadcasted to $\mathbb{R}^{n\times c}$ as the regularization term.

Normalization. Finally, we get the regularized grouping basis $\mathbf{R}$ by summing $\mathbf{R}^{raw}$ and $\mathbf{R}^{re}$ which is weighted with a learnable vector $\beta \in \mathbb{R}^{1\times 1\times n}$ initialized to 0.1. Then we apply softmax to obtain the soft grouping basis,

$$x_{i,j}^{*} = \frac{exp(x_{i,j})}{\sum_{i=1}^{n} exp(x_{i,j})} \tag{5}$$

where $j \in [1, c]$. Each $x_{i,j}^{*}$ in $\mathbf{R}^{*} \in \mathbb{R}^{n\times c}$ measures the possibility that the i-th channel of the feature map belongs to the j-th group.

Then we divide the feature map into n groups in channel dimension and normalize them respectively similarly to Eq. (8). The final output of CAN compute the weighted sum of the original input feature map and the normalized one as $CAN(\mathbf{X}) = \eta\bar{\mathbf{X}} + \mathbf{X}$, where η is a learnable scalar initialized as 0.

3.3 Networks and Loss Functions

The overall network architecture is shown in Fig. 3. Formally, let $\mathbf{I}$ be the ground truth, $\mathbf{G}$ and $\mathbf{H}$ denote its gradient and high-frequency residual map.

The task of the first stage is to generate a rough result with good structure information but insufficient texture details. Thus, we take the masked image $\hat{\mathbf{I}} = \mathbf{I} \odot \mathbf{M}$ as the input, corresponding gradient map $\hat{\mathbf{G}} = \mathbf{G} \odot \mathbf{M}$ ($\mathbf{M}$ as the image mask with value 1 for known region otherwise 0). Here, $\odot$ denotes the Hadamard product. The generator produces a rough result $\mathbf{I}_{rough} = G_1(\hat{\mathbf{I}}, \hat{\mathbf{G}}, \mathbf{M})$

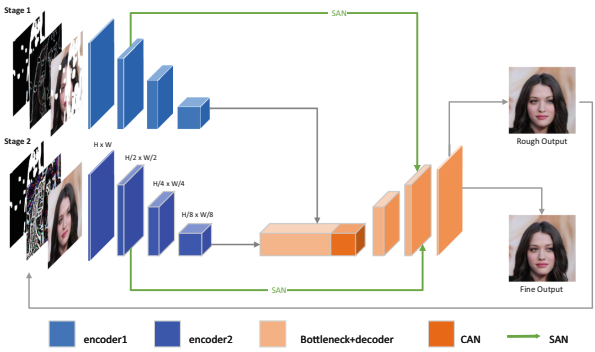

Fig. 3. The overview of our network architecture.

as part of the input of second stage, where G_1 represents the generator of the first stage, which is composed of enc_1 and dec.

The second-stage generator has its own independent encoder and a shared decoder with the first stage. Specifically, the masked high-frequency residual map $\hat{\mathbf{H}} = \mathbf{H} \odot \mathbf{M}$, instead of the gradient map, is another part of the input. In addition, smaller convolution kernels are used in the second stage's encoder than in the first stage's. We get the fine image $\mathbf{I}_{fine} = G_2(\hat{\mathbf{I}} + \mathbf{I}_{rough} \odot (1 - \mathbf{M}), \hat{\mathbf{H}}, \mathbf{M})$, where G_2 represents the generator of the second stage, which is composed of enc_2 and dec. The final predicted image is $\mathbf{I}_{pred} = \hat{\mathbf{I}} + \mathbf{I}_{fine} \odot (1 - \mathbf{M})$.

The Generator is trained over a joint loss, similar to the other methods, which consists of ℓ_1 loss ($\mathcal{L}_{\ell_1}$), perceptual loss ($\mathcal{L}_{perc}$), style loss ($\mathcal{L}_{style}$), and adversarial loss ($\mathcal{L}_G$). Besides, there is an orthogonal loss equal to the sum of $\mathcal{L}_{so}$ and $\mathcal{L}_{co}$ of two-stage, enabling attention map can learn various semantic attention. Therefore, the overall loss function of the proposed SCAN algorithm is as $\mathcal{L}_{total} = 10\mathcal{L}_o + \mathcal{L}_{\ell_1} + 0.1\mathcal{L}_{prec} + 250\mathcal{L}_{style} + 0.1\mathcal{L}_G$.

4 Experiments

All of the experiments in this paper are conducted in the dataset of CelebA-HQ and Paris Street View. The CelebA-HQ dataset is a high-quality version of CelebA that consists of 28000 train images and 2000 test images. The Paris Street View contains 14900 training images and 100 test images. We use Sobel filters to extract the gradient map and use the result after the image subtracting its Gaussian blur to get the high-frequency residual map. The irregular mask dataset used in this paper comes from the work of Liu [2]. We train the proposed model and the compared models on a single NVIDIA 2080Ti with a batch size of 8 until the generators converge, using Adam optimizer.

Table 1. The comparison of PSNR and SSIM over the CelebA-HQ and Paris Street View. Both quantitative evaluations are higher is better.

Dataset		CelebA-HQ			Paris street view		
Mask ratio		20%–30%	30%–40%	40%–50%	20%–30%	30%–40%	40%–50%
PSNR	EC [3]	27.18	25.29	23.33	27.28	25.80	24.10
	RN [11]	27.43	25.51	23.66	27.57	26.19	24.26
	Ours	**28.54**	**26.91**	**24.69**	**28.76**	**27.09**	**25.27**
SSIM	EC	0.917	0.902	0.862	0.875	0.851	0.804
	RN	0.929	0.912	0.871	0.887	0.864	0.812
	Ours	**0.953**	**0.933**	**0.894**	**0.918**	**0.873**	**0.840**

Table 2. The comparison of PSNR, SSIM, and MAE (Mean absolute error) over the Paris, in case of 30%–40% mask ratio. †Higher is better. △Lower is better.

	PC [2]	GC [10]	EC [3]	RN [11]	Ours
PSNR †	25.46	25.54	25.80	26.19	**27.09**
SSIM †	0.835	0.849	0.851	0.864	**0.873**
MAE (%) △	3.13	3.09	2.90	3.04	**2.53**

4.1 Quantitative Results

The Quantitative results in the test dataset of CelebA-HQ and Paris Street View are shown in Table 1, where some results produced by popular inpainting methods in comparison are also shown. In the case of different ratios of the damaged region, the table demonstrates the inpainting ability of our network, showing that our results are better than other results in PSNR (Peak Signal-to-Noise) and SSIM (Structural Similarity) metrics. In particular, Table 2 demonstrates our method has a more remarkable improvement in all metrics in detail when the mask ratio is 30%–40%. Compared with PC, GC, EC, RN, and RFR, our method can effectively use the correlation with the limited known regions information.

4.2 Qualitative Results

Figure 4 illustrates the visual inpainting results of different methods on the test set of CelebA-HQ and Paris Street View with mask ratios of 30%–40% and 40%–50%.

EC is misled by the edge information generated in the first stage when predicting the results. RN tends to produce smoother results and lacks texture details. Compared with EC and RN, the proposed method can better handle much larger damaged regions and achieves better subjective results, even houses with complex structural information on Paris Street View. Also, the visual

Fig. 4. Qualitative comparisons with EC and RN on Paris Street View and CelebA-HQ with the mask ratio of 30%–40% and 40%–50%.

comparison on CelebA-HQ shows obvious enhancement of our method, such as sharp facial contours, crisp eyes and ears, and reason-able object boundaries.

5 Conclusions

This paper proposed a two-stage image inpainting approach with a SCAN module. The SAN and CAN module normalize feature map in spatial and channel dimensions, respectively. Various experiments show that the proposed SCAN generates promising images and achieves the state-of-the-art performance. In addition, the CAN module improving the group normalization could also be generalized to similar image restoration tasks, including image denoising, conditional image generation, and image segmentation.

References

1. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
2. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 85–100 (2018)
3. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 (2019)
4. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536–2544 (2016)

5. Wang, N., Li, J., Zhang, L., Du, B.: Musical: multi-scale image contextual attention learning for inpainting. In: IJCAI, pp. 3748–3754 (2019)
6. Wang, Y., Chen, Y.C., Zhang, X., Sun, J., Jia, J.: Attentive normalization for conditional image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5094–5103 (2020)
7. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
8. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
9. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5505–5514 (2018)
10. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4471–4480 (2019)
11. Yu, T., et al.: Region normalization for image inpainting. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12733–12740 (2020)