On the use of attention in deep learning based denoising method for ancient Cham inscription images

Tien-Nam Nguyen, Jean-Christophe Burie, Le Thi Lan, Anne-Valérie Schweyer

On the use of attention in deep learning based denoising method for ancient Cham inscription images

Tien-Nam Nguyen[1](✉)[0000−0002−2984−697X], Jean-Christophe Burie[1][0000−0001−7323−2855], Thi-Lan Le[2][0000−0001−9541−3905], and Anne-Valerie Schweyer[4][0000−0002−1058−8835]

[1] Laboratoire Informatique Image Interaction (L3i) La Rochelle University , Avenue Michel Crépeau, 17042, La Rochelle Cedex 1, France
{tnguye28,jcburie}@univ-lr.fr
[2] School of Electronics and Telecommunications, Hanoi University of Science and Technology, Vietnam.
lan.lethi1@hust.edu.vn
[3] Centre Asie du Sud-Est (CASE), CNRS, Paris, France
anne-valerie.schweyer@cnrs.fr

Abstract. Image denoising is one of the most important steps in the document image analysis pipeline thanks to its good effect into the rest of the workflow. However, the noise in historical documents is totally different from the common noise present in other classical problems of image processing. It is particularly the case of the image of Cham inscriptions obtained by the stamping of ancient stele. In this paper, we leverage the advantage of deep learning to adapt with these noisy conditions. The proposed network follows an encoder-decoder structure by combining convolution/deconvolution operators with symmetrical skip connections and residual blocks for improving reconstructed image. Furthermore, global attention fusion is proposed to learn the relevant regions in the image. Our experiments demonstrate the proposed method can't only remove unwanted parts in the image, but also enhance the visual quality for the Cham inscriptions.

Keywords: Document Image Analysis · Historical Document · Image Denoising · Attention · Cham Inscription.

1 Introduction

Exploring cultural heritage has attracted many researchers these last decades. Historical handwritten documents are important evidence in order to understand historical events and especially the ones of extinct civilizations. The Cham inscriptions are written from the Cham language system, which has been used from the very early centuries AD in Champa (nowadays Vietnam coastal areas) and some nearby areas. The descendants of the Cham population represent one part of the community in Southeast Asia. Nowadays, the Cham inscriptions are

mainly carved on steles of stone. Over time, the aging and climatic conditions have damaged the characters and created bumps. Many unwanted parts or gaps appeared on the stones making the visual quality of the image degraded significantly. The readability has become a real challenge for archaeologists, historians, as well as for people curious about Cham culture. The preservation of this cultural heritage is an important problem that needs to be considered. Similar research works to preserve palm leaf manuscripts have been done in [3].

In Document Image Analysis (DIA), a binarization process is usually used as a pre-processing step before applying the Optical Character Recognition (OCR) step. However, as mentioned above, these inscriptions have been damaged. Moreover, the image of the inscriptions, we work with, are obtained by a stamping process which may also create noise. Thus, traditional image binarization can not work well on these types of inscriptions due to the various degradation. Hence, an adaptive approach to remove noise and improve the visual quality of these inscriptions is needed before analyzing the content.

With the tremendous performance in many computer vision tasks, deep neural networks (DNNs) have demonstrated the ability in not only traditional tasks but also on complex learning tasks. Despite the outcome from DNNs, applying the latest techniques to the document historical problems have not been studied enough yet. Image denoising task could be considered as an image translation task that maps one image from the noisy domain to the cleaned domain [17]. Inspired by this idea, the supervised [1, 2] or unsupervised [4, 5] approach have shown promising results in image denoising problems.

Attention is proposed as an auxiliary module to make model robust to salient information rather than learning insignificant background parts in the image. Attention has obtained successful results in natural language processing problems such as [20, 21] and is gradually used in computer vision problems [22]. Specifically, at different scales, an image contains different information. At lower scale it reflects general information while it contains detailed information at higher scale [30]. To efficiently leverage this information, we propose global attention fusion which accumulates attention from different scales to enhance denoised image quality.

Our main contributions are briefly detailed as follows: First, to the best of our knowledge, this is the first time attention module is embedded into image denoising method on historical documents. Second, we proposed an adaptive way to use the attention module in the training model and the global attention fusion module in cooperation with the loss function providing higher both qualitative and quantitative results. Finally, we tested the effectiveness of the proposed method by comparing it with different denoising methods of the literature on the Cham inscription dataset.

The rest of the paper is organized as follows. In section 2, we present a brief overview of related works on image denoising, especially for historical documents, based on both traditional and deep learning approaches. In Section 3, we introduced the Cham inscription dataset. In section 4, the details of the proposed approach are presented. The experimental results and comparison with tradi-

tional approaches are described in section 5. Finally, a conclusion with some future research directions are given in Section 6.

2 Related work

Image denoising. Usually, traditional methods have been applied for image denoising problems based on the characteristics of noise in the corrupted image. Such methods can be categorized as spatial filters [6–8] or transform domain [13, 14, 16]. The denoising in spatial domain is based on the observation that noise appears at high frequency and low pass filter is adapted to eliminate this noise. However, these filter-based methods are not robust as each type of noise requires distinct filter kernel size. Another approaches are based on the pioneer work of non-local means [9]. The denoised image is estimated by weighting with similar patches on different locations of image. To handle with the blurry results, some suitable regularization methods can be used to enhance the quality of the denoised image such as: total variation regularization [10], sparse prior [11], low rank prior [12]. In the transform domain, image will be converted to a new domain where the characteristics of the noisy part are different with the clean part. Fourier transform, Wavelet transform [13, 14, 16] are common transform domain approaches. The main disadvantage of transformation-based approach is choosing the kind of transform or wavelet bases which are suitable for data. Besides the model driven methods, data driven methods are conventional approaches which are based on statistical representation of the data. These methods try to represent a set of images into sub-components with some prior assumption then remove redundancy in the representation which is estimated the noise present in the image. Independent Component Analysis [34] and Principal Component Analysis [36] are widely represented for these methods. The BM3D algorithm [15] is an impressive work which leverages the advantage of spatial domain and the transform domain. In general, these traditional approaches depend on pre-defined rules, which reduce the robustness to different types of noises. Although these methods have demonstrated successful results, especially BM3D that achieves very promising results in comparison with deep learning based approaches, the main problem is that all the methods work under the assumption that noise is an additive white Gaussian noise with a given standard deviation. However, in real problems, this requirement is not adaptable, noise needs to be represented by a more complex function. Then most of the research shifted to deep learning based approaches. One of the first works using deep learning approach is proposed in [17]. They used a multi-layer perceptron (MLP) to learn the mapping directly from noisy to clean image. Based on this idea, many works have shown competitive results with very deep learning-based approaches such as: [1], [2], [37]. In the work of [1], they used a deep convolutional network made of two parts : an encoder for learning clean parts in the image and a decoder to reconstruct the original shape of the image from them. Instead of directly mapping from noisy image to noisy-free image, [2] proposed a model for leaning the noisy space in the image, the noisy-free image can be

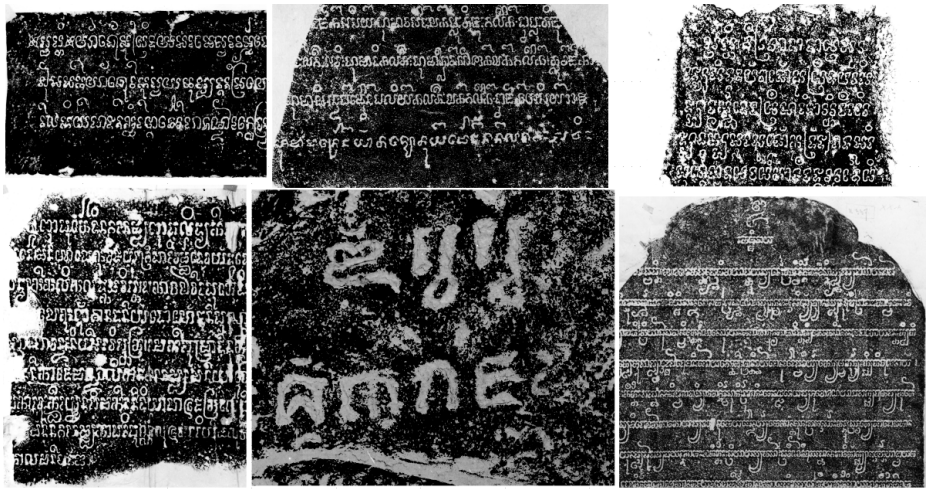

Fig. 1. Sample inscriptions in the Cham dataset

obtained by subtracting the noisy image and the noise extracted from the model. Related to our work, [18] and [19] applied generative modeling, successfully used in the context of historical handwritten document analysis. However, the mentioned approaches are equally considering the role of each pixel, the outcome results will not clearly distinguish between background (noise) and foreground pixels (characters). This hence lowers the qualitative and quantitative results. To resolve this issue, we adapted the attention module.

Attention. Based on the principle of human vision, attention mechanism is proposed to be robust to the important parts of the image instead of learning irrelevant parts by assigning higher weight to the useful region. Attention can directly be integrated as a component in the model. In general, attention can be split into two types: self-attention [23–25] which is computed from only one input feature and general attention [26] which is computed from two or more input features. Depending on the problems and the architecture of models, the appropriate attention is selected.

3 Dataset

We now present the Cham inscription dataset used in our research works. A preliminary introduction about Cham language can be found at [32]. Since no dataset and no ground truth were available, our first contribution was to build a dataset and the corresponding ground truth. The images of Cham inscriptions have been obtained by a stamping process. This work has been done in Vietnam by archaeologists during excavations in the field or by curators in museums, when the steles were deposited there. The stamping process allows to copy the inscriptions carved on the stone (ancient stele) on a large sheet of paper. How-

ever, all gaps and bumps are also "copied". The dimensions of the images are variable from 1000 to 6000 pixels either height or width. The sheets of paper are then scanned to obtain digital images. Our dataset, at the moment, consists of 100 "documents" (duplicates of ancient stone steles). Besides the limitation of available ground truth, the challenges of these inscriptions also come from the range of dating. The inscriptions that have been collected fall within a wide chronological range from the 6th to the 15th century AD and are written either in Cham or in Sanskrit. The use of each of these two languages leads to a slightly different writing system, which must be considered for the text analysis but they can be processed in the same way for the denoising task. In addition, the degradation of many inscriptions poses an additional challenge for working on Cham inscriptions. As we can see in Figure 1, there are many parts of the text that may seem undesirable. By simple observation, it is often impossible to tell whether certain strokes on the stone form part of a letter or a group of letters or are noise. The annotation has been done by a linguistic expert. A raster graphic editor has been used to remove all pixels of noise and correct the missing pixels of text. As the cleaning task is time-consuming and to avoid processing large-size documents, each "document" image has been split in text line images. Our dataset consists of 190 text line images that have been cleaned one by one by the expert. So, the ground truth consists of a set of binary images without any noise and with cleaned characters (according to the expert knowledge).

4 Proposed Approach

4.1 Architecture

As shown in Figure 2, the proposed model consists of two main components: a baseline encoder-decoder model and the attention module.

Baseline model: The baseline of the proposed model is inspired by Unet [27] that follows the encoder-decoder architecture. Input image is first processed by a consecutive Convolution-BatchNorm-Relu layer to the 1x1 size at the bottle layer. Then from the output of bottleneck layer, reconstructed image is achieved by doing the series Deconvolution-BatchNorm-Relu (up-sampling step). However, the up-sampling step is done from the bottleneck layer where the size of the data is small, much of the information will be lost in the reconstructed images. Instead of down-sampling the input image to the 1x1 size, the input image is down-sampled twice before going through a sequence of Resnet block [29] in the middle (the blue blocks in Figure 2). This type of architecture is similar to the coarse-to-fine generator in [28].

Attention: In order to reduce the redundancy of the information when using skip connection [1] between down sampling steps and up sampling steps, we adapt the attention gate introduced in [26]. This module selects appropriate information instead of keeping all information from the down-sampling step. The integration of the attention gate in skip connection can be explained as below. First, an attention map is generated from F_{D2} and F_{U2}. This attention map

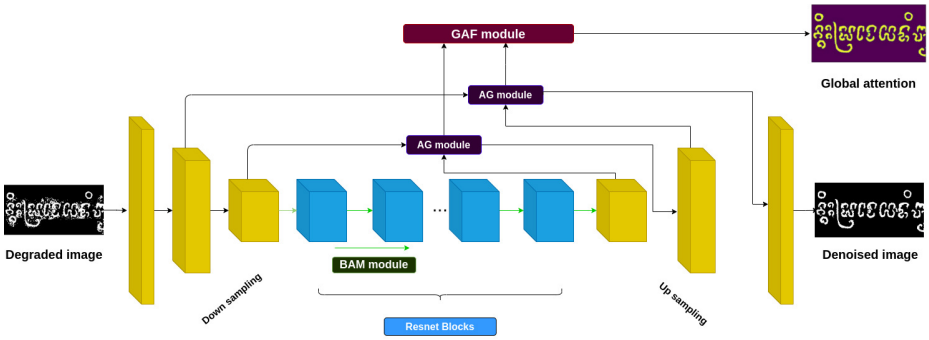

Fig. 2. Architecture of the proposed model.

has high value at the similar feature between F_{D2} and F_{U2}. The modification features F'_{D2} is computed through multiplying F_{D2} with the attention map.

$$F'_{D2} = F_{D2} * \sigma(W_n * (\omega((W_d * F_{D2} + W_u * F_{U2}))) \tag{1}$$

where F_{D2}, F_{U2} are the features at the down sampling step and up sampling step respectively, W_d, W_u and W_n are the parameters of convolution layers, ω is Relu activation function, σ is the sigmoid activation function. After that, we concatenate F'_{D2} and F_{U2} features as common skip connection then upsample it to F_{U1}.

$$F_{U1} = W_t * (concat([F_{U2}, F_{D2'}])) \tag{2}$$

where W_t are parameters of deconvolution layer. This process also repeats at the next upper scale. The details of module are represented in Figure 3.

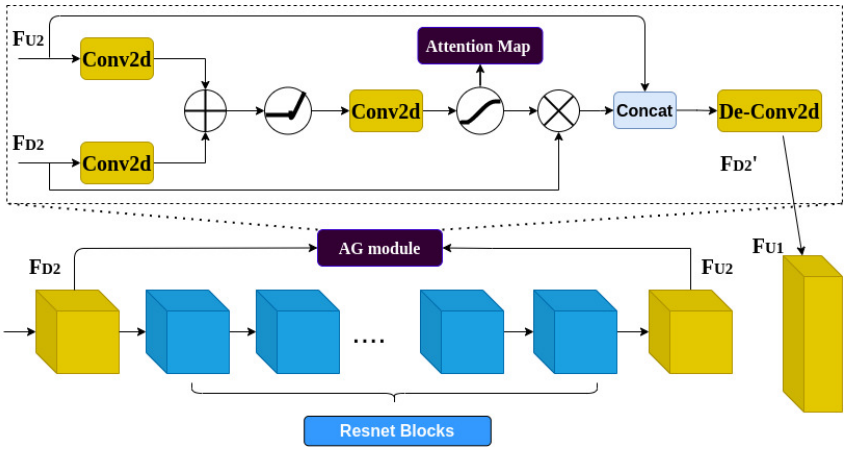

Fig. 3. Integration of Attention Gate in Skip Connection [26]

For the sequence Resnet block in the middle of our architecture, we modified the normal connection between Resnet block by adding the Bottleneck Attention Module (BAM) [24] after every Resnet block to improve the separation of the features at low-level such as: gaps or strokes on the surface and gradually focus on the exact target at a high-level of semantic (characters) by integrating attention from both channel and spatial information. After the sequence Resnet block + BAM (see in Figure 4), the model aims to highlight salient features of character regions while deducting the features of non-relevant regions. Shortly, refined features by BAM can be described as:

$$F' = F + F * \sigma(M_c(F) + M_s(F)) \tag{3}$$

where F is the output feature of the previous Resnet block. M_c, M_s are the channel attention and spatial attention, respectively. More details of each channel attention and spatial attention can be found in the [24].

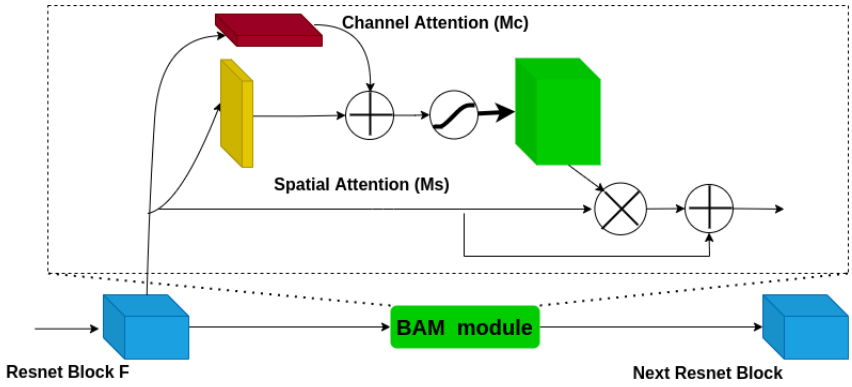

Fig. 4. Refined features by BAM [24]

Global attention fusion: Due to the difficulty when reading Cham inscriptions, it is more important to keep and enhance the quality of the characters than removing only the noise. Therefore, we proposed a global attention fusion module by integrating attention at multiple scales to help the model focus more on the pixels of the characters. This module works as below. First, we concatenate the attention map from different scales of the attention gate module.

$$C(A) = concat(A_1, resize(A_2)) \tag{4}$$

where A_i be the attention map generated from the attention gate at different scales, i is the scale of input image. The concatenated attention is then followed by a deconvolution layer to generate global attention map which has the same shape of input image then sigmoid activation function to normalize the coefficients of them to the range (0,1).

$$C_s = \sigma(W_A * C(A)) \tag{5}$$

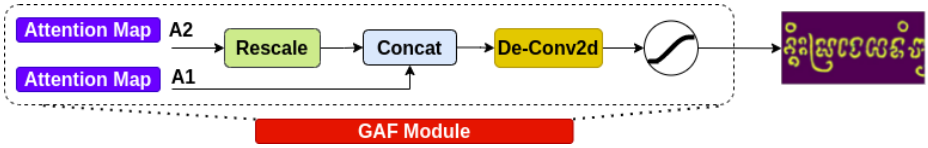

Fig. 5. Global Attention Fusion Module

where W_A is the parameters of deconvolution layer, σ is the sigmoid activation function. (See Fig. 5). These coefficients of global attention map (C_s) represent the confidence of the generated pixels for the character regions in the image. As we can see in Figure 6, at the beginning, only simple patterns which model easily to recognize are noises or characters, will correctly reconstructed with the high confidence score. After some epochs, this score gradually increases to reveal difficult patterns of the characters which have lower scores at the first epochs.

Fig. 6. Evolution of confidence score. Bright pixels have a high score, dark ones have a low score.

4.2 Objective Function

The training objective function combines two main functions. The first function, weighted L1 Loss (weight reconstruction loss), helps to reconstruct the denoised image of important regions closest to the ground truth images.

$$L_1^w = ||C_s * \hat{y} - y||_1 \tag{6}$$

where C_s is the confidence score output from GAF module. The second one is the perceptual loss L_p [30] which forces a generated image from the model to have a perceptual feature similar to the one of the ground truth images. Perceptual features are intermediate features extracted from different layers of a pre-trained model such as VGG, Inception or Resnet.

$$L_p = \sum_{i=1}^{n} \frac{1}{C_i * H_i * W_i} \sum_{h=1}^{H_i} \sum_{w=1}^{W_i} \sum_{c=1}^{C_i} ||F^i_{h,w,c}(\hat{y}) - F^i_{h,w,c}(y)||_1 \tag{7}$$

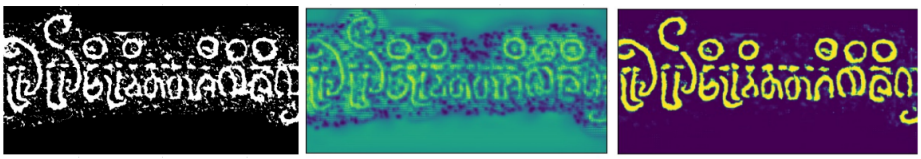

Fig. 7. Attention map with and without constraint on the C_s. We can see that in the middle of the figure, when the model is trained without constraint, character regions and empty regions have high coefficients. In the right image, when using constraint, the model aims to reveal only character regions, the other regions have a low score.

where $F^i_{h,w,c}$ is output feature from list n features named as 'relu1-2', 'relu2-2', 'relu3-3', 'relu4-3' of VGG network.
We found that if we used C_s in the equation 6 without constraint, the coefficients of C_s have high values in regions almost empty because for those regions, the model is easy to reconstruct. After that, the optimization for these regions will not change. (See Fig. 7). To avoid that we simply used the mean of C_s as an additional constraint that helps C_s to consider only the character regions instead of the empty regions. Overall the objective function can be described as:

$$L = \lambda_{l_1} * L_1^w + \lambda_{l_{cs}} * mean(C_s) + \lambda_{l_p} * L_p \tag{8}$$

where λ_{l_1}, $\lambda_{l_{cs}}$ and λ_{l_p} are the weights of the weight reconstruction loss (L_1^w), mean value final attention map (C_s) and perceptual loss (L_p) in the total loss, respectively. Each weight indicates the contribution of each component in the total loss function. The higher the weight is, the greater the contribution of the components is. The model has to balance both pixel level and high features level based on different weights of each loss.

5 Experiments

Dataset: We evaluated the proposed model on the dataset presented in section 3. This dataset consists of images of Cham inscriptions written in Cham or Sanskrit. Due to the different size of each inscription, we normalized the size of the training images. Instead of resizing the whole text line image which leads to distortion, we simply cropped the original image into sub-images which have a size of 256x512 pixels. 140 text line images were split into approximate 2500 images as training data while 50 other text line images were used as testing data. Furthermore, due to the limitation available data, augmentation strategy was also studied. We used some data augmentation below: random rotation (some text lines are not horizontal), random erasing, similar mosaic augmentation [31] (combining images by cutting parts from some regions and pasting them onto the augmented image).
Training Setting: We used the Adam optimizer with initial learning rate 0.0002. The weight λ_{l_1}, $\lambda_{l_{cs}}$ and λ_{l_p} were determined experimentally 1, 0.1,

2, respectively. Further studies will be done to evaluate the influence of these weights on the results but they are not in the scope of this paper. All models in our experiments were trained from scratch with the same the number of training data. The experiments were done on 2080Ti GPU 12GB memory.
Evaluation metrics: In order to evaluate the denoising performance, we used two metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM). PSNR is a quality metric which measures how the denoised image is close to the ground truth image while SSIM metric measures the similarity structure between two images. The highest PSNR and SSIM values indicate the better results.

5.1 Ablation Study

Table 1. Ablation study on the sub-component in our model

Model	Avg-PSNR	Avg-SSIM
Baseline[27]	15.45 ± 2.65	0.898 ± 0.050
Baseline[27] + AG[26]	15.57 ± 2.62	0.900 ± 0.049
Baseline[27] + BAM[24]	15.86 ± 2.76	0.904 ± 0.048
Proposed model	**15.98 ± 2.82**	**0.905 ± 0.048**

Table 1 shows the ablation study on the effects of each component in the proposed model which goes from baseline model, BAM module, AG module, and the global attention fusion module. On the SSIM metric, the results are similar because on text line images, the structure of the image is simple, and all methods can simply preserve the original structure. On the PSNR metric, we can see that the use of the BAM module and AG module have slightly improved results. The proposed method can achieve better results than each module combined separately and shows an improvement in comparison to the baseline model.

5.2 Comparison with different approaches

Table 2 presents a comparison of our approach with other methods of the literature applied on our dataset. We split these methods into two main groups. In the first group, the input image is directly processed without any training step. The second group consists of methods with a training step integrating knowledge from both original (degraded image) and clean image. The first group gathers traditional denoising methods (TV [10], NLM [9], BM3D [15]) and binarization methods (Otsu [38], Sauvola [40], Niblack [39]). The traditional methods, in average, PSNR is improved but the value of the SSIM metric is lower than the original images. This can be explained by the fact that these methods are efficient on regions where the size of the noise is relatively small, but it leads the output image blurrier, the SSIM score is lower than the original input image. The results with binarization methods show that these approaches are not adapted since they don't reduce the noise (lower PSNR than original image) even if they

Table 2. Quantitative results on our dataset

Method	Avg-PSNR	Avg-SSIM
Original	11.93	0.651
TV [10]	12.68	0.567
NLM [9]	12.01	0.489
BM3D [15]	11.64	0.623
Ostu [38]	10.96	0.713
Sauvola [40]	10.44	0.689
Niblack [39]	10.59	0.694
NMF [35]	12.63	0.749
FastICA [34]	12.69	0.754
Unet [27]	15.54	0.900
Pix2pix [33]	14.05	0.875
Pix2pixHD [28]	13.26	0.855
Proposed method	**15.98**	**0.905**

enhance the global visual quality of the image providing a better SSIM than original image. Qualitative results on images with different methods on the low and high degradation are shown respectively in Figure **??** and Figure 9.

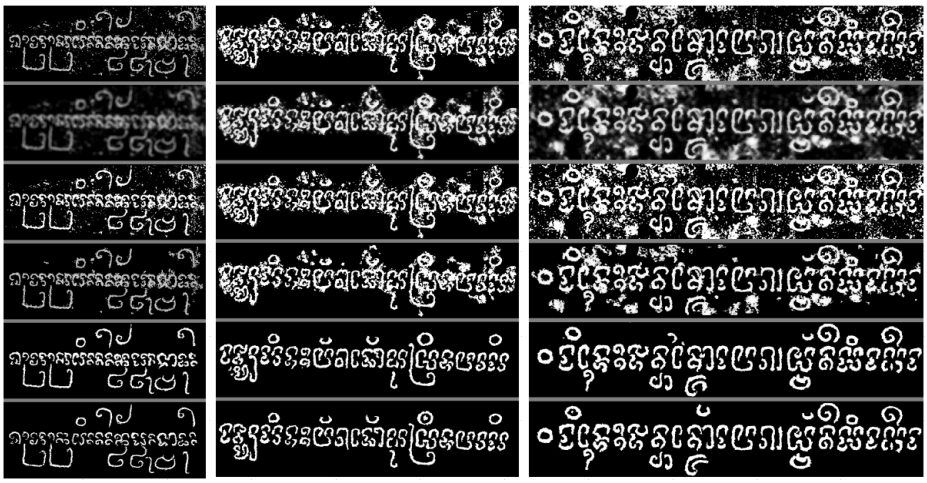

Fig. 8. Some qualitative results on Cham inscriptions with **low** degradation. From top to bottom: Original image, NLM, Otsu, FastICA, Proposed method, Ground truth.

In the second group, both NMF [35] and FastICA [34] methods improve both the PSNR and SSIM metrics but we observed that the denoising ability is very limited when the patterns of noise are similar to some parts of the characters affecting the readability. The last methods of this second group consist of

deep learning-based approaches. All methods have significantly boosted both the quantitative (PSNR and SSIM metrics) and qualitative results. The proposed method achieves better results on both metrics. Besides enhancing denoising results compared to the other methods, our approach gave better results because it generates pixels with a high confidence value in the foreground and thus provides a better visual quality.

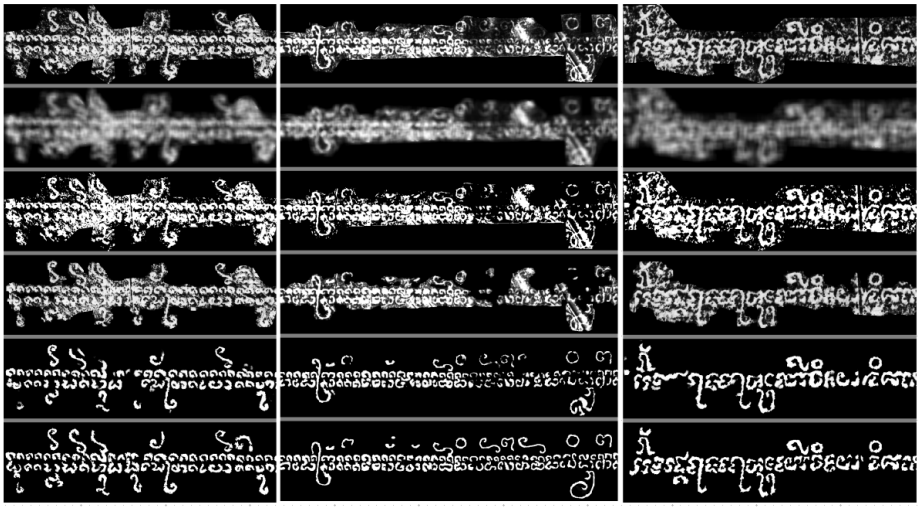

Fig. 9. Some qualitative results on Cham inscriptions with **high** degradation. From top to bottom: Original image, NLM, Otsu, FastICA, Proposed method, Ground truth.

5.3 Qualitative Results

If the proposed method improves statistically the quality of the image, the readability is also an important parameter for the next step : the recognition process. In order to evaluate the relevance of the approach we asked an expert in Cham language to estimate the performance of our method. We asked her to analyse qualitatively two different criteria : performance on noise removal and character readability. For each criterion, we defined four level assessments: very bad, bad, normal, and good, corresponding respectively to the 1, 2, 3, 4 score. The qualitative evaluation is obtained by computing, for each experiment detailed in table 1, the average score for each criterion. However, instead of evaluating the results on the whole testing set, it was split into three subsets depending of the historical period : 7th-9th century (5 images), 10th-12th century (19 images) and 13th-15th century (26 images) The quality of the inscriptions depends on their age due to the fact that the damages are more important on the older stones.

The separation of images into chronological categories has been motivated by the evolution of writing, starting from often irregular and less codified scripts, passing through the blossoming and mastery of characters and ending with less refined scripts and less distinguished characters, creating more confusion for deciphering. So, we created 3 categories by this separation. Table 3 presents the qualitative results. For the noise removal criterion, we observed that the integration of AG[26] or BAM[24] into the baseline model improves significantly the results compared to the simple baseline model. For character readability criterion, the proposed method outperforms the other approaches. This qualitative evaluation shows the effect of the GAF module by encouraging the model to generate character pixel with higher confidence value. The improvement of contrast between foreground and background part as well as the quality of the characters, it increases the readability of the Cham inscription.

Table 3. Average score of the qualitative evaluation on our testing dataset split in 3 sets depending of the historical period

Aspects	Model	7th-9th	10th-12th	13th-15th
Noise Removal	Baseline[27]	3 ± 0	2.45 ± 0.60	1.53 ± 0.58
	Baseline[27] + AG[26]	3 ± 0	2.6 ± 0.50	1.85 ± 0.61
	Baseline[27] + BAM[24]	3 ± 0	2.8 ± 0.52	2.31 ± 0.68
	Proposed method	**3.4** ±0.55	**2.84** ± 0.66	**2.65** ± 0.77
Character Readability	Baseline[27]	2.2 ± 0.45	1.94 ± 0.41	1.34 ± 0.41
	Baseline[27] + AG[26]	3.2 ± 0.45	2.15 ± 0.50	1.96 ± 0.41
	Baseline[27] + BAM[24]	3 ± 0.70	2.05 ± 0.40	1.88 ± 0.41
	Proposed method	**3.4** ± 0.55	**2.52** ± 0.77	**2.34** ±0.41

6 Conclusion

In this work, we present an approach based on attention model for denoising old Cham inscription images. We have first introduced a new dataset for the image denoising problem of Cham inscriptions. We have detailed several experiments and analysed the benefits as well the disadvantages of each method. The quantitative and qualitative evaluations show that the proposed approach improves the quality of the Cham inscription in terms of noise removal and character readability. However, a room of improvement is certainly possible by using more data for the training step in order to consider more styles of degradation. So, we envisage to increase the size of the dataset to improve our model. In future work, we also plan to tackle the next step : the recognition of Cham Inscriptions.

7 Acknowledgment

This work is supported by the French National Research Agency (ANR) in the framework of the ChAMDOC Project, n°ANR-19-CE27-0018-02.

References

1. Mao, X.J., Shen, C., Yang, Y.B.: Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. Proceedings of the 30th Int. Conf. on Neural Information Processing Systems, 2016
2. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Trans. on image processing, 3142–3155 (2017)
3. Kesiman, M.W.A., Valy, D., Burie, J.C., Paulus, E., Suryani, M., Hadi, S., Verleysen, M., Chhun, S., Ogier, J.M.: Benchmarking of document image analysis tasks for palm leaf manuscripts from southeast asia. Journal of Imaging 4(2) (2018)
4. Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., Aila, T.: Noise2noise: Learning image restoration without clean data. Int. Conf. on Machine Learning, (2018)
5. Krull, A., Buchholz, T.O., Jug, F.: Noise2void-learning denoising from single noisy images. In: IEEE Conf. on Computer Vision and Pattern Recognition. (2019)
6. Pitas, I., Venetsanopoulos, A.N.: Nonlinear digital filters: principles and applications, vol. 84. Springer Science Business Media (2013)
7. Wiener, N.: Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications. MIT press Cambridge (1950)
8. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Sixth Int. Conf. on computer vision. pp. 839–846. IEEE (1998)
9. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: IEEE Conf. on Computer Vision and Pattern Recognition. pp. 60–65 (2005)
10. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena 60(1-4), 259–268 (1992)
11. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. IEEE Trans. on Image processing pp. 3736–3745 (2006)
12. Dong, W., Shi, G., Li, X.: Nonlocal image restoration with bilateral variance estimation: a low-rank approach. IEEE Trans. on image processing pp. 700–711 (2012)
13. Choi, H., Baraniuk, R.: Analysis of wavelet-domain wiener filters. In: Proceedings of the IEEE-SP Int. Symposium on Time-Frequency and Time-Scale Analysis. pp. 613–616 (1998)
14. Ram, I., Elad, M., Cohen, I.: Generalized tree-based wavelet transform. IEEE Trans. on Signal Processing 59(9), 4199–4209 (2011)
15. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Trans. on image processing 16(8), 2080–2095 (2007)
16. Portilla, J., Strela, V., Wainwright, M.J., Simoncelli, E.P.: Image denoising using scale mixtures of gaussians in the wavelet domain. IEEE Trans. on Image processing 12(11), 1338–1351 (2003)
17. Burger, H.C., Schuler, C.J., Harmeling, S.: Image denoising: Can plain neural networks compete with bm3d? In: 2012 IEEE Conf. on Computer Vision and Pattern Recognition. pp. 2392–2399 (2012)
18. Dumpala, V., Kurupathi, S.R., Bukhari, S.S., Dengel, A.: Removal of historical document degradations using conditional gans. In: ICPRAM (2019)
19. Souibgui, M.A., Kessentini, Y.: De-gan: A conditional generative adversarial network for document enhancement. IEEE Trans. on PAMI (2020)
20. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conf. on Empirical Methods in Natural Language Processing. pp. 1412–1421. Lisbon, Portugal (Sep 2015)

21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is All you Need. In Proceedings of the 31th Int. Conf. on Neural Information Processing Systems, 2017.
22. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conf. on Computer Vision. pp. 213–229. Springer (2020)
23. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conf. on computer vision. pp. 3–19 (2018)
24. Park, J., Woo, S., Lee, J.Y., Kweon, I.S.: Bam: Bottleneck attention module. British Machine Vision Conference, 2018
25. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: Int. Conf. on machine learning. pp. 7354–7363. PMLR (2019)
26. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D.: Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis 53, 197–207 (2019)
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Int. Conf. on Medical image computing and computer assisted intervention. pp. 234–241. Springer (2015)
28. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings IEEE Conf. on computer vision and pattern recognition. (2018)
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conf. on computer vision and pattern recognition. pp. 770–778 (2016)
30. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conf. on computer vision. pp. 694–711 (2016)
31. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
32. Nguyen, M.T., Shweyer, A.V., Le, T.L., Tran, T.H., Vu, H.: Preliminary results on ancient cham glyph recognition from cham inscription images. In: 2019 Int. Conf. on Multimedia Analysis and Pattern Recognition (MAPR). pp. 1–6. IEEE (2019)
33. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conf. on computer vision and pattern recognition. pp. 1125–1134 (2017)
34. Hyvarinen, A., Hoyer, P., Oja, E.: Sparse code shrinkage: Denoising by nonlinear maximum likelihood estimation. Advances in Neural Information Processing Systems pp. 473–479 (1999)
35. Févotte, C., and Idier, J.: Algorithms for nonnegative matrix factorization with the beta-divergence. Neural computation 23(9), 2421–2456 (2011)
36. Deledalle, C.A., Salmon, J., Dalalyan, A.S., et al.: Image denoising with patch based pca: local versus global. In: BMVC. vol. 81, pp. 425–455 (2011)
37. Zhang, K., Zuo, W., Zhang, L.: Ffdnet: Toward a fast and flexible solution for cnnbased image denoising. IEEE Trans. on Image Processing pp. 4608–4622 (2018)
38. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. on systems, man, and cybernetics 9(1), 62–66 (1979)
39. Niblack, W.: An Introduction to Digital Image Processing. Strandberg Publishing Company, DNK (1985)
40. Sauvola, J., Pietikainen, M.: Adaptive document image binarization. Pattern recognition 33(2), 225–236 (2000)