Exploiting Aliasing for Manga Restoration

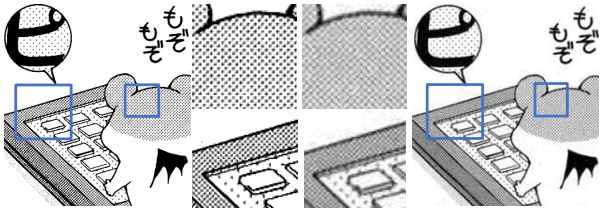
Minshan Xie[1,2,*] Menghan Xia[1,*] Tien-Tsin Wong[1,2,†]
[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2] Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology, SIAT, CAS
{msxie, mhxia, ttwong}@cse.cuhk.edu.hk

Abstract

As a popular entertainment art form, manga enriches the line drawings details with bitonal screentones. However, manga resources over the Internet usually show screentone artifacts because of inappropriate scanning/rescaling resolution. In this paper, we propose an innovative two-stage method to restore quality bitonal manga from degraded ones. Our key observation is that the aliasing induced by downsampling bitonal screentones can be utilized as informative clues to infer the original resolution and screentones. First, we predict the target resolution from the degraded manga via the Scale Estimation Network (SE-Net) with spatial voting scheme. Then, at the target resolution, we restore the region-wise bitonal screentones via the Manga Restoration Network (MR-Net) discriminatively, depending on the degradation degree. Specifically, the original screentones are directly restored in pattern-identifiable regions, and visually plausible screentones are synthesized in pattern-agnostic regions. Quantitative evaluation on synthetic data and visual assessment on real-world cases illustrate the effectiveness of our method.

1. Introduction

Manga, also known as Japanese comics, is a popular entertainment art form. One of the key differences between manga and other illustration types is the use of *screentones*, regular or stochastic black-and-white patterns to render intensity, textures and shadings (Figure 1(a)). Although furnishing impressive visual impact, the existence of screentones makes it tricky to resample manga images. For instance, when being undersampled, the bitonal regular patterns may get ruined and present incoherent visual effects (Figure 1(c)). Unfortunately, it is common to see such screentone artifacts from the manga images over the Internet (e.g. Manga109 [16]), probably due to the poor scanners or storage limitation in the old days. In this background, we are motivated to restore these low-quality legacy mangas and show their original appearances.



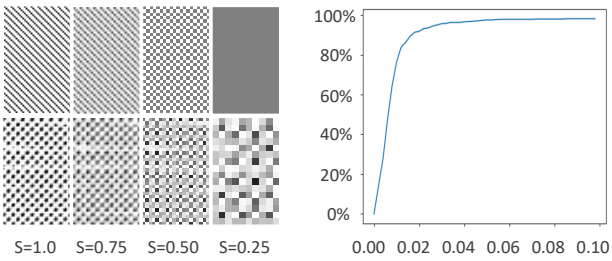
(a) Standard manga (b) Blow-up (c) Degraded manga

Figure 1: The screentones in the manga image with insufficient resolution are blurry while the desired screentones should be sharply bitonal. The image comes from the Manga109 dataset [16]. Akuhamu ©Arai Satoshi

Unlike natural images dominating with low-frequency components, manga images mainly consist of regular high-frequency patterns that are pickier at the representing resolution. Specifically, for a quality bitonal manga image of resolution T, it is generally impossible to present the screentones in a both bitonal and perceptually consistent manner on the resolution $S \neq T_k \in \{kT|k = 1, 2, 3, ..., n\}$. That means, to restore a manga image, we first need to figure out a target resolution that is able to present the potential target screentones, and then restore the screentones from the degraded ones at that scale. Apparently, this tricky requirement excludes the feasibility of existing Single Image Super-Resolution (SISR) methods [3, 5, 8] and image restoration methods [15, 30, 2, 19]. Instead, our key idea is inspired by an interesting observation that the aliasing caused by downsampling the bitonal screentones is usually distinctive on the downscaling factor and screentone type, as illustrated in Figure 2. These may serve as informative clues to infer the original screentones and their associated resolution.

To this end, we propose an innovative two-stage manga restoration method. First, we utilize the Scale Estimation Network (SE-Net) to predict the target resolution from the degraded screentones. There are usually multiple screen-

*Equal contributions.
†Corresponding author.



(a) Screentone downscaling (b) Correlation statistics

Figure 2: Our observation. (a) Aliasing from screentone downscaling is visually distinctive, depending on the scale factor and screentone type. (b) Statistic of scale prediction error on a synthetic dataset (degraded scale ranges $1.0 \sim 4.0$). A prediction error below 0.02 is achieved over 91% samples, indicating the strong correlation between aliasing property and the applied downscaling factors.

tones within a single manga image, and each of them may contribute differently to the prediction accuracy. This is effectively tackled through our proposed spatial voting scheme based on confidence. At the predicted resolution scale, we restore the region-wise bitonal screentones via the Manga Restoration Network (MR-Net). Considering the different degradation degrees, the manga image is restored with two parallel branches: the original screentones are restored for pattern-identifiable regions, and random screentones are synthesized under intensity conformity for pattern-agnostic regions. Specifically, this region-wise classification is determined adaptively through a learned confidence map. Separately, the two networks are trained over the mixture dataset of synthetic manga and real ones, in a semi-supervised manner.

We have evaluated our method on our synthetic testset and some real-world cases. Quantitative evaluation demonstrates that our restored manga images achieve high PSNR and SSIM on synthetic data. Meanwhile, qualitative evaluation of real-world cases evidences the potential for practical usage. In summary, this paper has the contributions:

- The first manga restoration method that restores the bitonal screentones at a learned resolution.
- A manga restoration network that restores the region-wise screentones adaptively based on a learned confidence map.

While our current method is tailored for the manga restoration problem, our proposed framework has the potential to be extended to natural images containing the regular textures. For example, the checkerboard artifact resulted from undersampling the regular textures should share a similar property as the screentones.

2. Related Work

2.1. Manga Screening

Attempts have been made to generate screentone manga automatically from grayscale/color images or line drawings. Qu et al. [18] applied a variety of screentones to segments based on the similarity between texture, color, and tone to preserve the visual richness. However, the method failed to restore bitonal manga from the degraded version as the degraded screentones maybe significantly differ from the original patches. Li et al. [13] presented an effective way to *synthesize* screen-rich manga from line drawings. Tsubota et al.[22] synthesize manga images by generating pixel-wise screentone class labels and further laying the corresponding screentones from database. However, these methods are highly dependent on the screentone set and cannot generate the original bitonal screentones. In contrast, our method attempts to recover the original version of the degraded manga by learning the degradation rules of screentones with generalization to real-world cases.

2.2. Single Image Super-Resolution

As a classic vision task, Single Image Super-Resolution (SISR) aims at reconstructing the high-resolution (HR) version from the low-resolution (LR) images. Traditional methods mainly leverage dictionary learning [29] or database retrieval[1, 4, 21] to reconstruct the high-frequency details for the low-resolution input. However, due to the limited representation capability of hand-crafted features and lack of semantic level interpretation, these methods struggle to achieve photorealistic results.

Recently, as deep learning techniques on the rise, the state-of-the-art SISR has been updated continuously by these data-driven approaches. Given pairs of LR and HR images, some studies [3] attempt to solve it as a regression problem that maps LR images to their corresponding HR images. Many follow-ups reached a more accurate HR image by designing better network architectures, such as VDSR [9], SRResNet[12], LapSRN [11], or more powerful loss functions, like EDSR[14]. However, these methods tend to generate blurry results as they failed to recover the lost high-frequency signal that has little correlation with the LR image. To recover these lost details, some approaches [12, 5, 24] adopt generative adversarial networks (GANs) [6] to generate stochastic details. SRGAN [12] attempts to recover the lost details by adopting a discriminator to tell what kind of high-frequency details look natural. As a step further, a super-resolution method of arbitrary scale [8] is proposed to reconstruct the HR image with continuous scale factor, which is the most related work to our method. However, all these methods, mainly working on natural images, never consider the scale suitability when recovering the high-resolution images. This is just the

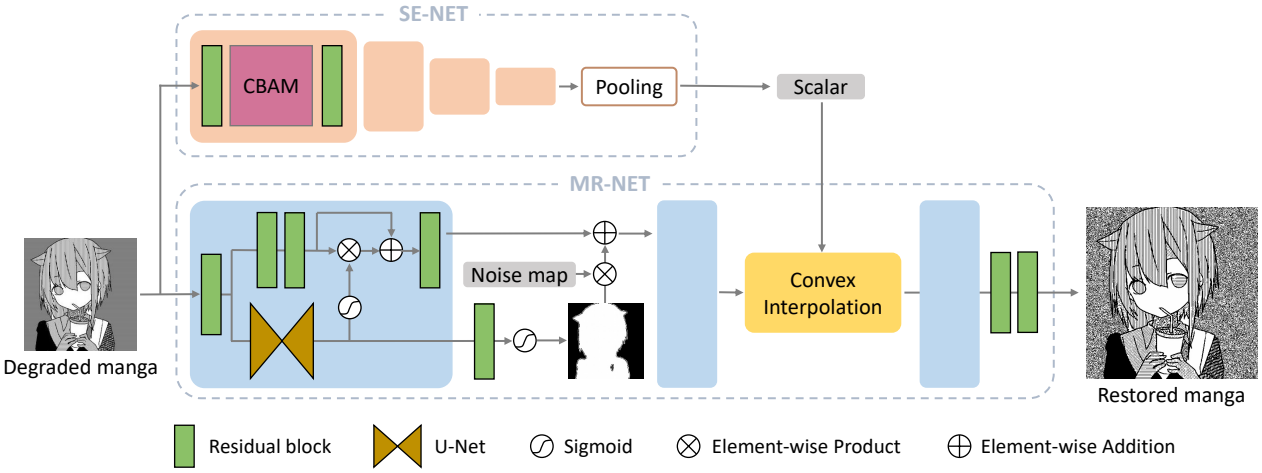


Figure 3: System overview. Given the degraded manga image, the SE-Net first estimates the scalar that is required to upscale, and then the MR-Net restores the manga image at the predicted resolution scale.

intrinsic difference between SISR and our problem. Indeed, our method attempts to first obtain a proper resolution from the degraded screentones which then helps to restore the bitonal nature.

3. Scale-Awared Manga Restoration

Given a degraded manga image, we aim to restore the bitonal screentones to be as conformable as possible to the original version. As illustrated in Figure 3, it is formulated by a two-stage restoration framework including restorative scale estimation and manga screentone restoration. The detailed network architectures are provided in the supplementary material.

3.1. Problem Formulation

Let $\mathbf{I}_{\text{gt}}$ be the original bitonal manga image. Generally, the degraded image $\mathbf{I}_x$ can be modeled as the output of the following degradation:

$$\mathbf{I}_x = (\mathbf{I}_{\text{gt}} \otimes \kappa) \downarrow s_{\text{gt}} + \mathbf{N}_\varsigma, \tag{1}$$

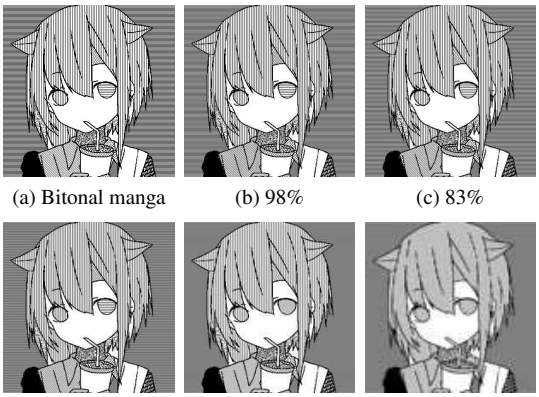
where $\{\kappa, s_{\text{gt}}, \varsigma\}$ parameterizes the degradation process. $\mathbf{I}_{\text{gt}} \otimes \kappa$ denotes the convolution between a blur kernel κ and the image $\mathbf{I}_{\text{gt}}$, $\downarrow s_{\text{gt}}$ denotes the downsampling operation with the scale factor s_{gt}. Without losing generality, $\mathbf{N}_\varsigma$ describes other potential noises induced by scanning process or lossy image format like JPEG, which overall is modelled as additive noises with standard deviation ς.

This seems very similar to the formula adopted in super-resolution [26], but the problem is crucially different due to the special nature of manga restoration. As introduced in Section 1, it is impossible to recover the original bitonal screentones from degraded ones unless it is represented with appropriate resolution. To tackle this problem, we need to: (i) figure out the target resolution by estimating the desired scale factor: $s_y = g(\mathbf{I}_x) \rightarrow s_{\text{gt}}$; (ii) perform the manga screentone restoration at the estimated resolution, which is a comprehensive deconvolution and denoising process: $\mathbf{I}_y = f(\mathbf{I}_x, s_y) \rightarrow \mathbf{I}_{\text{gt}}$. It makes sense because of the distinctive correlation between $\mathbf{I}_{\text{gt}}$ and $\mathbf{I}_x$ that conditions on s_{gt} and the screentone type, as observed in Figure 2. In particular, we try to model the functions $g(\cdot)$ and $f(\cdot)$ by training two neural networks respectively.

3.2. Restorative Scale Estimation

Based on the degraded manga image, we utilize the Scale Estimation Network (SE-Net) to estimate the downscaling scalar that has been applied to the original bitonal manga image. This is a prerequisite of the subsequent manga restoration that requires a screentone-dependent restorative resolution.

Scale Estimation Network (SE-Net). The SE-Net takes the degraded image $\mathbf{I}_x$ as input and outputs the estimated scale factor s_y for further restoration. Figure 3 shows the abstract structure of the SE-Net, which cascades four downsample modules and an adaptive pooling layer. As a common case, a single manga image contains multiple screentone regions and each of them degrades to some different extent depending on the screentone pattern types, as shown in Figure 4. Consequently, these screentone regions might be informative differently to infer the downscaling factor s_{gt}, which motivates us to adopt the Convolutional Block Attention Module (CBAM)[27] to focus on deterministic regions and ignore noisy ambiguous regions. Since the attention of CBAM is performed in the feature domain along both channel and spatial dimensions, the intermediate feature maps are adaptively optimized with sufficient flexibility along with these downsample modules.



(a) Bitonal manga (b) 98% (c) 83%

(d) 67% (e) 50% (f) 25%

Figure 4: Manga degradation with different downscaling factors. Different screentones will have different degradations with the same downscaling factor. The screentones on background are degraded to plain region which gives no clue for restoration with 50% resolution while the screentones on foreground still retains informative patterns.

Besides, to get the scalar output, we perform the adaptive pooling on the feature maps, where a global spatial pooling and a fully-connected layer are combined.

Loss Function. The SE-Net is trained with the loss function comprised of two terms: scale loss $\mathcal{L}_{\text{scl}}$ and consistency loss $\mathcal{L}_{\text{cons}}$.

$$\mathcal{L}_{\text{SE}} = \mathcal{L}_{\text{scl}} + \alpha\mathcal{L}_{\text{cons}}, \tag{2}$$

where $\alpha = 0.1$ balances the magnitude of the two terms.

Scale loss. Given degraded image $\mathbf{I}_x$, the scale loss $\mathcal{L}_{\text{scl}}$ is to encourage the SE-Net to generate a scale factor s_y which is as close as possible to the ground truth s_{gt}.

$$\mathcal{L}_{\text{scl}} = \|s_y - s_{\text{gt}}\|_1, \tag{3}$$

Consistency loss. When only trained on synthetic data based on $\mathcal{L}_{\text{scl}}$, we find that it cannot generalize well to real-world cases. For example, on a scanned manga book, the predicted scale factors for different images from the same volume or even for different patches from the same image can be substantially different. Thus, we further introduce a consistency loss $\mathcal{L}_{\text{cons}}$ to enforce the SE-Net to generate a consistent scale factor for the patches from the same manga image. Actually, this loss term benefits in two aspects: on the one hand, it stabilizes the network training by further introducing extra supervision; on the other, it enables semi-supervised training on the mixture data of synthetic manga and real-world manga and thus promotes the generalization performance on real-world cases.

$$\mathcal{L}_{\text{cons}} = \|s_y - \frac{1}{M}\sum_{i=1}^{M} s_y^i\|_1, \tag{4}$$

where s_y^i denotes the predicted scale factor from the i-th of M patches cropped from the same degraded image $\mathbf{I}_x$.

3.3. Discriminative Restoration

Based on the estimated scale factor, we utilize the Manga Restoration Network (MR-Net) to restore the screentones for the target manga image. According to the screentone degradation degrees, the MR-Net restores the manga image discriminatively on different regions: reconstruct the original screentones for pattern-identifiable regions while synthesizing plausible screentones for pattern-agnostic regions.

Manga Restoration Network (MR-Net). The MR-Net takes the degraded manga image $\mathbf{I}_x$ and the desired scale factor s_{gt} as input, while output the confidence map $\mathbf{M}_c$ and the restored manga image $\mathbf{I}_y$. Figure 3 shows the abstract structure of the MR-Net, which employs the Residual Attention Module (RAM) [23] as backbone unit to capture the screentone clue and restore the screentone regularity. Specifically, the attention features of the first RAM are further transformed to a single-channel confidence map $\mathbf{M}_c$ that is used to selectively introduce noises to the feature maps. The intuition is that the output manga image will be generated through two paths implicitly, i.e. reconstruction path and synthesis path, and the random noises are injected to add external variation for the regions under the charge of the synthesis path. The second RAM further prepare the features for spatial upsampling, which is implemented by the convex interpolation block [20] with learned neighborhood interpolative coefficients. Specifically, we interpolate a target pixel from N known neighboring pixels $\{p_1, p_2, ..., p_N\}$ by computing the weighted sum: $\sum_{i=1}^{N} \alpha_i p_i$, where $\sum_{i=1}^{N} \alpha_i = 1$ and $\forall \alpha_i \geq 0$. Then, the upsampled feature maps are transformed to the restored manga image by the rest layers.

Loss Function. The optimization objective of the MR-Net comprises five terms: pixel loss $\mathcal{L}_{\text{pix}}$, confidence loss $\mathcal{L}_{\text{conf}}$, binarization loss $\mathcal{L}_{\text{bin}}$, intensity loss $\mathcal{L}_{\text{itn}}$ and homogeneity loss $\mathcal{L}_{\text{hom}}$, written as:

$$\mathcal{L}_{\text{MR}} = \mathcal{L}_{\text{pix}} + \phi\mathcal{L}_{\text{conf}} + \omega\mathcal{L}_{\text{bin}} + \kappa\mathcal{L}_{\text{itn}} + \gamma\mathcal{L}_{\text{hom}}, \tag{5}$$

The empirical coefficients $\phi = 0.5$, $\omega = 0.5$, $\kappa = 0.5$ and $\gamma = 0.02$ are used in our experiment.

Pixel loss. The pixel loss $\mathcal{L}_{\text{pix}}$ ensures the restored manga image $\mathbf{I}_y$ to be as similar as possible with the ground truth $\mathbf{I}_{\text{gt}}$ on those pattern-identifiable regions and helps the network to reconstruct the original bitonal image. Here, we measure their similarity with the Mean Absolute Error (MAE), as defined in

$$\mathcal{L}_{\text{pix}} = \|\mathbf{M}_c \odot |\mathbf{I}_y - \mathbf{I}_{\text{gt}}|\|_1, \tag{6}$$

where $\odot$ denotes element-wise multiplication and $|\cdot|$ denotes the operation to take the element-wise absolute value. The loss attention mechanism avoids overfitting to low-confidence regions, potentially focusing less on ambiguous regions.

Confidence loss. The confidence loss $\mathcal{L}_{\text{conf}}$ encourages the model to extract as many pattern-identifiable regions as possible. Sometimes, it is quite ambiguous to visually detect whether certain screentone degradation is pattern-identifiable or not. Instead, we formulate it as a confidence map $\mathbf{M}_c$ that is learned adaptively. Based on the prior that most degraded screentones are restorable, we encourage the model to restore as much as possible screentones through

$$\mathcal{L}_{\text{conf}} = 1.0 - \|\mathbf{M}_c\|_1. \tag{7}$$

Here, the confidence map $\mathbf{M}_c$ has 1 represent pattern-identifiable regions and 0 indicates pattern-agnostic regions.

Binarization loss. To generate manga with bitonal screentones, we introduce the binarization loss $\mathcal{L}_{\text{bin}}$ to encourage the network to generate black-and-white pixels, which is defined as

$$\mathcal{L}_{\text{bin}} = \||\mathbf{I}_y - 0.5| - 0.5\|_1. \tag{8}$$

Intensity loss. The intensity loss $\mathcal{L}_{\text{itn}}$ ensures that the generated manga image $\mathbf{I}_y$ visually conforms to the intensity of the target image $\mathbf{I}_{\text{gt}}$. According to the low-frequency pass filter nature of Human Visual System (HVS), we compute this loss as:

$$\mathcal{L}_{\text{itn}} = \|\text{G}(\mathbf{I}_y) - \text{G}(\mathbf{I}_{\text{gt}})\|_1, \tag{9}$$

where G is a Gaussian blur operator with the kernel size of 11×11. Specially, when calculating the intensity loss on the real-world cases, we resize the input to the target resolution and further smooth it with a Gaussian filter, which is still qualified guidance to constrain the intensity similarity. In practice, this loss benefits in two folds. For the ambiguous regions that are insufficient to restore the original screentones, we can still leverage this loss term to generate screentones with similar tonal intensity. In addition, it allows the training on real-world manga data to promote generalization performance.

Homogeneity loss. The screentone homogeneity loss aims to impose the homogeneity within each screentone region. With the previous loss, we observe that the restored manga images sometimes have inconsistent screentones even in the same homogeneous regions. To alleviate this problem, we encourage the screentone features within each region to be similar through the homogeneity loss $\mathcal{L}_{\text{hom}}$. Here, we measure the screentone difference in the domain of ScreenVAE map [28] that represents the screentone pattern as a smooth and interpolatable 4D vector and enables the pixelwise metrics (e.g. MSE) to be effective. In particular, we formulate the homogeneity loss $\mathcal{L}_{\text{hom}}$ as:

$$\mathcal{L}_{\text{hom}} = \frac{1}{N}\sum_{i=1}^{N} \|\text{SP}_i(\Phi(\mathbf{I}_y)) - \mu(\text{SP}_i(\Phi(\mathbf{I}_y)))\|_2, \tag{10}$$

where $\Phi(\cdot)$ extract the ScreenVAE map of a manga image, SP_i denotes the the i-th superpixel that is achieved by segmenting on the ground truth manga $\mathbf{I}_{\text{gt}}$, and $\mu(\cdot)$ computes the mean value.

4. Experimental Results

4.1. Implementation details

Data Preparation. Manga109 [16] is a public manga dataset, containing a total of about 20000 pieces of degraded image. However, the resolution of the manga images is low. Currently, there is no high-resolution public manga dataset which we can directly use as ground truth. Fortunately, Li et al.[13] proposed an effective manga synthesis method, which fills in line drawings with diverse screentones. To prepare paired training data, we synthesized 3000 pieces of bitonal manga images with the resolution of 2048×1536, and simulate various degradation through the random combination of: (i) downsampling with multiple scale factors $s \in [1.0, 4.0]$; (ii) JPEG compression with different quality factors $q \in [50, 100]$; (iii) Gaussian noise with varied standard deviation $\mathcal{N}(0.0, 5.0 \sim 15.0)$.

Training Scheme. To favor the model generalization on real-world cases, we apply a semi-supervised strategy to train the SE-Net and the MR-Net separately, i.e. both paired synthetic data and unpaired real-world data are used for training. In particular, for the synthetic data, all the losses are used, i.e. $\mathcal{L}_{\text{scl}}$ and $\mathcal{L}_{\text{cons}}$ in the first stage, and $\mathcal{L}_{\text{pix}}$, $\mathcal{L}_{\text{conf}}$, $\mathcal{L}_{\text{bin}}$, $\mathcal{L}_{\text{itn}}$, and $\mathcal{L}_{\text{hom}}$ in the second stage. For the Manga109 which has no ground truth available, only $\mathcal{L}_{\text{cons}}$ (stage 1) and $\mathcal{L}_{\text{conf}}$, $\mathcal{L}_{\text{bin}}$, $\mathcal{L}_{\text{itn}}$ (stage 2) are used.

We trained the model using PyTorch framework [17] and trained on Nvidia TITANX GPUs. The network weights are randomly initialized using the method of [7]. During training, the models are optimized by Adam solver [10] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initialized to 0.0001.

4.2. Scale Estimation Accuracy

We evaluate the accuracy of our Scale Estimation Network (SE-Net) on synthetic data with different rescaling ranges. As tabulated in Table 1, the accuracy decreases as the scale factor increases, because lower-resolution generally means severe degradation and hence involves more ambiguity for information inference. The average accuracy of the whole range $[1, 4]$ is 0.9896. In other words,

Table 1: Accuracy evaluation of the estimated scale factor on synthetic data.

Methods	Upscaling Range	Estimation Accuracy
w/o CBAM	[1,2]	0.9550
	(2,3]	0.9538
	(3,4]	0.9514
	[1,4]	0.9542
w CBAM	[1,2]	0.9823
	(2,3]	0.9936
	(3,4]	0.9929
	[1,4]	0.9896

Table 2: Scale prediction on the images from the same volume that shares the same ground-truth scale factor.

	Synthetic data ($s_{\text{gt}} = 2$)		Real data	
	$\mu(s_{\text{y}})$	$\sigma(s_{\text{y}})$	$\mu(s_{\text{y}})$	$\sigma(s_{\text{y}})$
w/o $\mathcal{L}_{\text{cons}}$	2.0483	0.1893	1.7995	0.4247
w $\mathcal{L}_{\text{cons}}$	2.0472	0.1217	1.2845	0.1346

for a degraded image with downscaling factor of T, our estimated scale factor is expected to be $(1 \pm 0.0104)T$. In addition, we study the effectiveness of the CBAM and our consistency loss respectively. We construct baseline module of the CBAM by removing the attention block, resulting in a simple residual block. We can observe that the CBAM improves the performance obviously, since the attention mechanism facilitates the network to focus on the informative regions while ignoring ambiguous regions.

Besides, we explore the role of the consistency loss $\mathcal{L}_{\text{cons}}$, which is mainly motivated to stabilize the training and generalize to real-world manga data. As the result shown in Table 2, it makes a significant improvement in the prediction stability but negligible accuracy gain on synthetic data. This is because the scale loss $\mathcal{L}_{\text{scl}}$ can guarantee a sufficiently high accuracy already. In contrast, it indeed causes a stable numerical result on real-world data.

4.3. Manga Restoration Quality

Comparison with Baselines. After obtaining a proper scale factor, we can restore the manga image through the Manga Restoration Network (MR-Net). To evaluate the performance, we compare it with three typical super-resolution approaches: EDSR[14] which is a regression-based method, SRGAN[12] which is a GAN-based method, and MetaSR[8] which is a method of arbitrary scale. The first two methods are trained with our dataset with a given scale factor ($\times 2$). MetaSR and our MR-Net are trained with scale factors ranged in $[1, 4]$. As most of screentones in our synthetic data lose their regularity when $\times 3$ downscaled, we evaluate the performance on a synthetic dataset with the scale factors ranged in $[1, 3]$. On those degraded images with scale factors $T \neq 2$, the evaluation on EDSR[14] and SRGAN[12] is performed by rescaling their results to the expected resolution. In particular, to avoid reference ambiguity, we quantitatively evaluate the restoration quality only on pattern-identifiable regions, as shown in Figure 5.

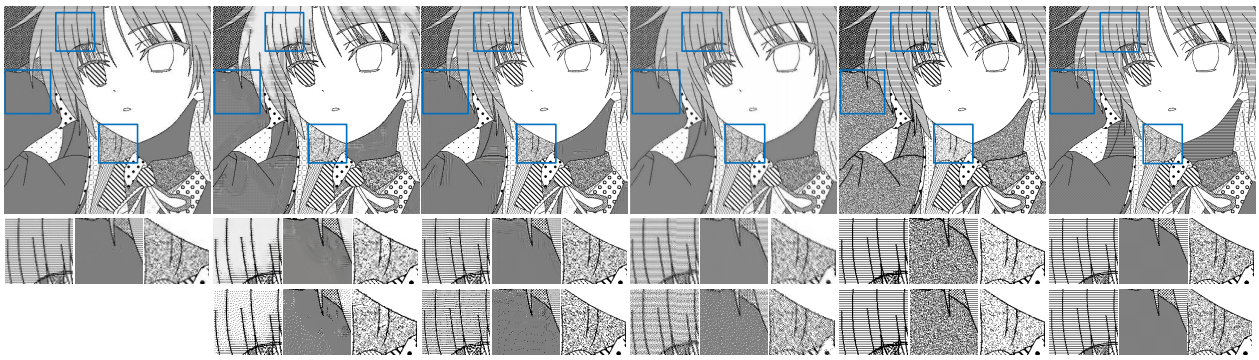
(a) Degraded manga (b) Confidence map (c) Ours (d) Grouth truth

Figure 5: Manga restoration results with synthetic data. Some ambiguous regions are degraded into a plain regions which has no clue to restore the original version.

We report experiment results in Table 3 using PSNR, SSIM[25] and SVAE. SVAE evaluates the screentone similarity between the generated results and the ground truth. It is achieved by comparing the ScreenVAE map [28] which is a continuous and interpolative representation for screentones. We can find that our model outperforms EDSR[14], SRGAN[12] and MetaSR[8] when the scale factor is various. Anyhow, our method can not achieve superiority over SRGAN[12] at scale factor of 2 when SRGAN is trained to handle exactly the $\times 2$ scale while our model is trained for various target scales. However, as mentioned earlier, the model trained with fixed scale is infeasible to solve our problem in practical scenarios. When combined with our SE-Net, MetaSR[8] can be roughly regarded as a reasonable baseline of our MR-Net. Note that our key concept is the scale-aware manga restoration, and the comparison with MetaSR that is provided with the ground-truth scale factors, is just to verify the structure effectiveness of the MR-Net. The quantitative results shown in Table 3 illustrates the superiority of our MR-Net structure that adopts a flexible attention mechanism and discriminative restoration strategy.

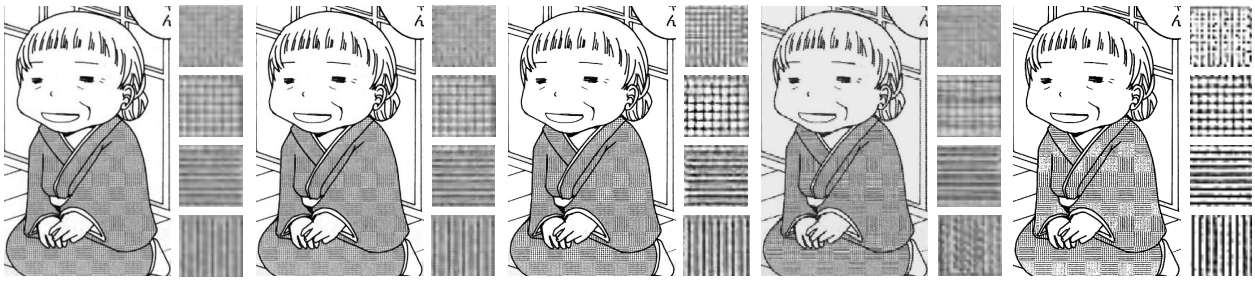
Figure 6 compares the visual results on typical synthetic examples. Our method successfully restores the bitonal manga image from the degraded screentones. For the region where the information has totally lost after resampling, our result generates random but consistent bitonal screentones, leading to better visual results. Meanwhile, our screentones are consistent over regions and can be directly binarized with little information loss.

Evaluation on Real-world Cases To validate the generalization, we test our method on some real-world cases (Manga109[16]). Results show that our method can restore visually pleasant results with clear screentones from the real-world degraded manga at the estimated upscaling resolution, as shown in Figure 7. As one may observe,



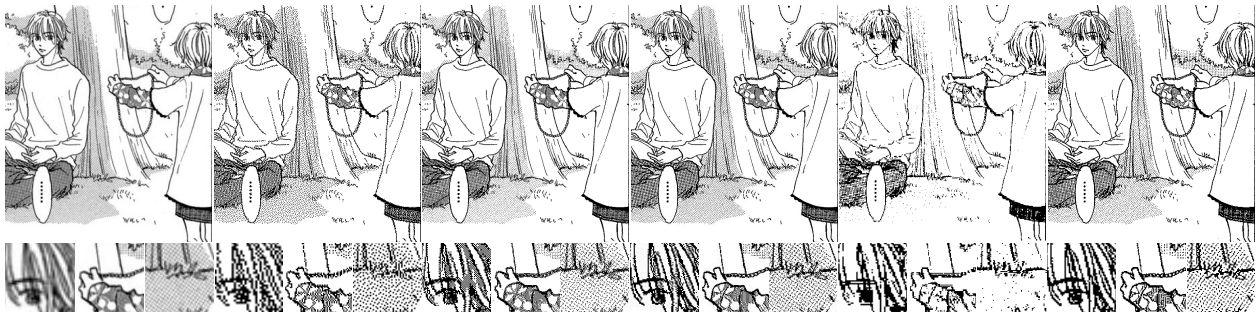
(a) Degraded manga (b) EDSR[14] (c) SRGAN[12] (d) MetaSR[8] (e) Ours (f) Grouth truth

Figure 6: Manga restoration results for synthetic data. Binarized results are shown on the bottom. EDSR[14], SRGAN[12] and MetaSR[8] may generate blurry screentones while our method can restore the bitonal nature.



(a) Degraded manga (b) EDSR[14](200%) (c) SRGAN[12](200%) (d) MetaSR[8](127%) (e) Ours(127%)

Figure 7: Manga restoration results for real-world case. The screentones are shown on the right. The image comes from the Manga109 [16]. Akuhamu ©Arai Satoshi

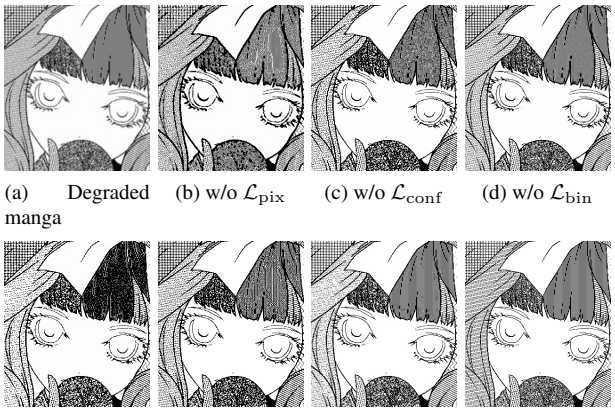


(a) Degraded manga (b) Bitonal manga (c) EDSR[14] (d) SRGAN[12] (e) MetaSR[8] (f) Ours

Figure 8: Manga restoration results for real world case with bitonal nature. (b) is the binarized result under original resolution. (c) and (d) are restored under 200% resolution while (e) and (f) are resotored under 150% resolution. The image comes from the Manga109 [16]. HaruichibanNoFukukoro ©Yamada Uduki

Table 3: Restoration accuracy of pattern-identifiable regions.

Resolution	$s_{\mathrm{gt}} = 2$			$s_{\mathrm{gt}} \in (1, 2]$			$s_{\mathrm{gt}} \in (2, 3]$		
Metric	PSNR(↑)	SSIM(↑)	SVAE(↓)	PSNR	SSIM	SVAE	PSNR	SSIM	SVAE
EDSR[14]	13.1695	0.6992	0.0318	13.9010	0.6206	0.0734	9.3550	0.2615	0.0717
SRGAN[12]	**14.8810**	**0.7829**	**0.0183**	14.8987	0.8132	0.0353	**12.5510**	0.5418	0.0527
MetaSR[8]	10.029	0.2385	0.1006	12.3722	0.4032	0.0779	8.1153	0.1149	0.1011
Ours	11.5547	0.7101	0.0255	**16.8054**	**0.8485**	**0.0222**	12.0333	**0.6214**	**0.0415**



(a) Degraded manga (b) w/o $\mathcal{L}_{\text{pix}}$ (c) w/o $\mathcal{L}_{\text{conf}}$ (d) w/o $\mathcal{L}_{\text{bin}}$

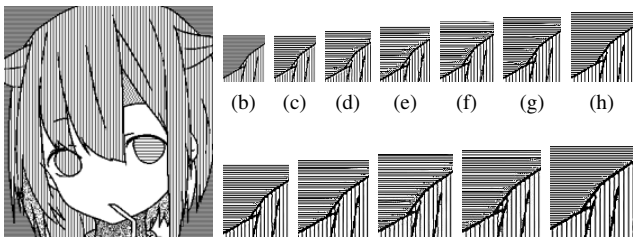
(e) w/o $\mathcal{L}_{\text{itn}}$ (f) w/o $\mathcal{L}_{\text{hom}}$ (g) Ours (h) Target

Figure 9: Ablation study for individual loss term.

our method can restore better results even with smaller resolutions. Since EDSR[14] and SRGAN[12] are trained with specific scale factors, they may not restore the periodic information for some unseen screentones. MetaSR[8] failed to restore the bitonal nature. Our method is also friendly to do binarization, as shown in Figure 8. We can see that although the regularity can be visually restored by EDSR[14] and SRGAN[12] under a larger scale factor, the results cannot be directly binarized which may destroy the structures. In contrast, our method can generate consistent screentones without destroying the structures.

Ablation Study for Individual Loss Terms. To verify the effectiveness of individual loss terms, we conduct ablation studies by visually comparing the generated output of different trained models without individual loss terms (Figure 9). The pixel loss $\mathcal{L}_{\text{pix}}$ is the essential component to guarantee to restore the original image. Without the intensity loss $\mathcal{L}_{\text{itn}}$, the pattern-agnostic regions may not follow the intensity constraint and thus generate undesired screentones. Meanwhile, the homogeneity loss $\mathcal{L}_{\text{itn}}$ is important for generating consistent screentones in the pattern-agnostic regions. In comparison, the combined loss can help the network to generate bitonal and consistent screentones for degraded manga images (Figure 9 (g)).

Robustness to Restoration Scale. We argue that manga restoration requires to conduct at an appropriate resolution because of the target screentone is bitonal and usually regular. When the restorative resolution is not matched, the restored screentones may either cause blurry grayscale intensity or present irregular patterns. To verify the necessity of the scale estimation, we study the performance of the MR-Net with different resolutions. As shown in Figure 10, only the results with ground-truth resolution (Figure 10 (h)) achieve visually pleasant bitonal screentone patterns.



(a) Degraded manga (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m)

Figure 10: Manga restoration results under different resolutions. (b) is the degraded patch and (c)-(m) are the restored patches under resolutions ranges $100\% \sim 200\%$.

4.4. Limitation and Discussion

Our method still suffers from some limitations. Our model may fail to restore the bitonal screentones for some real-world cases. This is related to several aspects. Firstly, there are still gaps between synthetic data and real-world cases. Although our method improves the generalization in a semi-supervised manner, we may still fail to generalize to some unseen patterns. Secondly, in real-world applications, some degraded manga images are degraded by multiple times and have some other unconsidered operations, which are beyond the assumption of our problem setting.

The pattern-agnostic regions are restored with screentones only under the condition of intensity constraint, which may cause perceptual inconsistency with the contextual screentones. In our future works, we will try to generate controllable screentone types with user input. Xie et al.[28] proposed an effective point-wise representation of screentones, called ScreenVAE map. We may provide the ScreenVAE value as a hint for the pattern-agnostic regions and constrain the generated screentones to have similar ScreenVAE values, along with intensity constraint.

5. Conclusion

In this paper, we propose a deep learning method for manga restoration with learned scale factor. Our method first predicts a suitable scale factor for the low-resolution manga image. With the predicted scale factor, we further restore the high-resolution image which has bitonal and homogeneous screentones. Our method achieves high accuracy on synthetic data and can generate plausible results on real-world cases.

Acknowledgement

This project is supported by Shenzhen Science and Technology Program (No.JCYJ20180507182410327) and The Science and Technology Plan Project of Guangzhou (No.201704020141).

References

[1] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004. 2

[2] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 576–584, 2015. 1

[3] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1, 2

[4] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002. 2

[5] Weifeng Ge, Bingchen Gong, and Yizhou Yu. Image super-resolution via deterministic-stochastic synthesis and local statistical rectification. In *SIGGRAPH Asia 2018 Technical Papers*, page 260. ACM, 2018. 1, 2

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 5

[8] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1575–1584, 2019. 1, 2, 6, 7, 8

[9] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 2

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[11] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2599–2613, 2018. 2

[12] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2, 6, 7, 8

[13] Chengze Li, Xueting Liu, and Tien-Tsin Wong. Deep extraction of manga structural lines. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 2, 5

[14] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2, 6, 7, 8

[15] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *Advances in Neural Information Processing Systems*, pages 1673–1682, 2018. 1

[16] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 1, 5, 6, 7

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[18] Yingge Qu, Wai-Man Pang, Tien-Tsin Wong, and Pheng-Ann Heng. Richness-preserving manga screening. *ACM Transactions on Graphics (SIGGRAPH Asia 2008 issue)*, 27(5):155:1–155:8, December 2008. 2

[19] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018. 1

[20] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020. 4

[21] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 1920–1927, 2013. 2

[22] Koki Tsubota, Daiki Ikami, and Kiyoharu Aizawa. Synthesis of screentone patterns of manga characters. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 212–2123. IEEE, 2019. 2

[23] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 4

[24] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2

[25] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[26] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[27] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3

[28] Minshan Xie, Chengze Li, Xueting Liu, and Tien-Tsin Wong. Manga filling style conversion with screentone variational autoencoder. *ACM Transactions on Graphics (SIGGRAPH Asia 2020 issue)*, 39(6):226:1–226:15, December 2020. 5, 6, 8

[29] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 2

[30] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017. 1