Neurocomputing 331 (2019) 199–212

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Multi-scale semantic image inpainting with residual learning and GAN

Libin Jiao, Hao Wu*, Haodi Wang, Rongfang Bie*

College of Information Science and Technology, Beijing Normal University, Beijing 100875, China

ARTICLE INFO

Article history:
Received 2 April 2018
Revised 25 October 2018
Accepted 13 November 2018
Available online 23 November 2018

Communicated by Dr. Haijun Zhang

Keywords:
Semantic inpainting
Image seamless fusion
Residual learning
Generative adversarial net

ABSTRACT

Image inpainting aims to fill the corrupt area semantically and recover the semantic and detailed information; however, concurrent methods suffer from convergence crash and arbitrarily paste the inference of the missing area into the corrupt context. In this paper, we study a combination of an encoder–decoder generator for image semantic inpainting and a multi-layer convolutional net for image seamless fusion, which is capable of restoring image effectively and seamlessly. Specifically, the encoder–decoder generator learns and extracts the latent compressed representations of missing areas from the context of a corrupt image, and further predicts a semantically correct estimation of the missing area based on the latent representations. The consecutive convolutional net smooths the discrepancy between the original image context and the estimation and seamlessly merges predictions and original images. The skip connections between the encoder and the decoder bridge the backward propagation of gradients, therefore boost the learning ability of the generator and stabilize the convergence of reconstruction loss. The performance and superiority of our method are illustrated and demonstrated on the real-world dataset qualitatively and quantitatively, and the experiments manifest acceptable semantic inpainting results, which significantly illustrates the effectiveness of our model.

1. Introduction

Rendering missing regions in a corrupt image by a non-semantically inpainting method is involved since the ill-posed image restoration problem fails in recovering semantic information and finer high-frequency details. Pixel-wise inpainting methods can seldom recognize the semantic entities, which leads to fixing the missing region with totally irrelevant patches; the inpainting result, therefore, hardly satisfies the perception of human beings. On behalf of image semantic inpainting method using deep learning design, Context encoder [1] practically tackles the semantic prediction of the missing region; it, however, can hardly stabilize its output after being trained on a limited dataset and through limited epochs. Moreover, the discontinuous boundaries between the original image and the inference region distract the human perceptual attention on the evaluation of the recovery of semantic information; thus, the hallucination across the areas notoriously decreases the reliability of the recovery for the perception of human beings. The “Copy-and-Paste” strategy is commonly used in existing inpainting methods due to its intuitiveness and convenience; however, without the image fusion processing, the coherence of the recovered image typically degrades, and the perceptually quality also remains unsatisfied after inpainting. Thus, concurrent inpainting methods can be typically promoted by fixing the aforementioned drawbacks.

Recently some breakthroughs in the architecture of convolutional networks principally thrive in the application circumstances on account of the accuracy and speed promotion of image classification, and these strategies effectively improve the quality of multi-scale and deep feature extraction. Starting with AlexNet [3], well-customized convolutional networks popularized themselves in Computer Vision, including ZFNet [4], GoogLeNet [5], VGGNet [6], and ResNet [7]; they dramatically outperformed the runner-up in the image recognition task due to their striking representative feature extraction. Residual skip connections [7] learn deep and shallow features by reusing the features passed through identity maps and combine them into underlying and desired feature maps. Bypassing the intermediate nonlinear activations, residual skip connections propagate gradients to minimize the loss when the network is being optimized and enable front layers to be principally calibrated, which assists in converging the network. As a practical implementation in image recognition, ResNet [7] thoroughly extended the depth of canonical counterparts, up to 152 layers, and won the competition of ILSVRC 2015. Inspired by the above residual blocks, a multi-layer convolutional net is employed to smooth the discontinuous boundaries between the generated region and

* Corresponding authors.
E-mail addresses: 92xianshen@mail.bnu.edu.cn (L. Jiao), wuhao@bnu.edu.cn (H. Wu), whd@mail.bnu.edu.cn (H. Wang), rfbie@bnu.edu.cn (R. Bie).

https://doi.org/10.1016/j.neucom.2018.11.045
0925-2312/

the original context, and the fusion coherency will be effectively restored.

In this paper, we discuss an improved model for restoring corrupt images semantically and seamlessly, which is based on an encoder–decoder architecture with residual skip connections and a multilayer convolutional net. Our model aims to estimate the missing regions in a corrupt image and can restore the semantic details correctly. Specifically, the encoder–decoder generator learns and extracts latent compressed representations of the missing area from a context of a corrupt image, and it further predicts a semantically correct estimation of these missing areas based on the latent representations. Subsequently, the consecutive convolutional net smooths the discrepancy between the original image context and the estimation, and it seamlessly incorporates prediction and original images. The skip connections between the encoder and decoder of generator bridge the backward propagation of gradients, and therefore, boost the learning ability of the generator and stabilize the convergence of holistic inpainting loss. Eventually, the concatenation of these two components semantically inpaints the missing hole and smooths the discontinuous boundaries when the prediction and the left context are assembled.

The main contributions of this paper are concluded as follows:

1. We customize an end-to-end model for semantic and seamless image inpainting, which concatenates an encoder–decoder generator with residual learning to inpaint semantically and a multi-layer convolutional net for image seamless fusion. The outputs of our model can effectively estimate the missing regions of a corrupt image.
2. We also manifest the assistance of skip connections in model designing, which benefits learning semantic representation from the context of an image and accelerates the convergence when training.
3. With fully-convolutional design, extending our model to multi-scale image inpainting with fine-tuning is unchallenging. We test our model on the dataset with size 256×256, and the model achieves acceptable results in the experiment.

The paper is organized as follows: Section 2 summarizes some related work regarding image inpainting, generation, and manipulation in recent years. Section 3 introduces our fully-convolutional image inpainting architecture, including the framework overview, our content network design, the joint loss function, and brief implementation details. Experiments and results are presented in Section 4 in which the evaluations regarding the human perceptual and the quantitative comparison, the ablation study of components, the robustness of adversarial loss are performed, and the results of the multi-scale image inpainting test are in this section. Section 5 discusses our experiments and concludes the paper.

2. Related work

We summarize some related work regarding exemplar-based or neural image inpainting methods, including structure & texture reconstruction, generative adversarial nets, semantic image inpainting & manipulation, and image detail improvements.

Structure and texture reconstruction using convolutional neural networks: Recently convolutional neural networks (ConvNet, CNN) thrived in the computer vision fields, including image classification [3,5–10], semantic segmentation [11–13] and structure prediction [14–19], and the hierarchical deep features were extracted by customized convolutional structures, such as the multi-layer smaller receptive fields covering more pixels [6], multi-scale convolutional kernels embedded in Inception module [5], and residual blocks learning residual features [7]. Aside from achieving great performance on semantic image classification, instance-level classification tasks, and pixel-level classification tasks [20], ConvNets were extended to fulfilling image prediction tasks, such as structure prediction, texture prediction and image generation [21–31], which created new neural applications widely.

Generative Adversarial Nets (GAN) and variants: Generative adversarial net (GAN) [32] aimed to formulate a parameterized model fitting the latent distribution of natural images within the given training set. Consisting of a generator for producing fake samples and a discriminator for differentiating fake and real samples, GAN learned the latent distribution without any supervision and could both generate new samples and classify samples with merely fine-tuning. Variants of GAN were proposed to enrich the application of GAN. Vanilla GAN [32] focused on unsupervised learning: the model should be capable of estimating the latent distribution on the dataset without supervised labels, while some variants tended to manually control the generated samples: conditional GAN [33] augmented the prior knowledge of the samples, and InfoGAN [34] aimed to find the connection between the features of generated samples and the corresponding controlling knobs. The drawback of lacking pair-wise training samples in the generation has been considered, and some trial solutions were proposed, such as Cycle GAN [35], Dual GAN [36], Disco GAN[37]. DCGAN [38] was an essential solution that focused on training a stable GAN structure in which both the generator and the discriminator could generate and differentiate real and fake samples respectively. Additionally, GAN can also facilitate clothing segmentation, retrieval [39], and saliency detection [40]. Recently, theoretical researches came to attention to improve the generating and differentiating performance, such as Wasserstein GAN [41,42] based on the Wasserstein distance, its improved implementation WGAN-GP [43] with gradient penalty, LSGAN [44] based on least square distance, LS-GAN, and GLS-GAN based on Lipschitz continuity [45].

Semantic image inpainting and manipulation: Semantic inpainting focuses on holistic content recovering by minimizing the predefined reconstruction loss (L2 loss). By minimizing the reconstruction loss and adversarial loss when training the ConvNet model, Context encoder [1] filled the missing hole in the corrupt image and adopted the training losses to control the distortion of generated images: the reconstruction loss reflected the holistic difference between the estimation and its corresponding ground truth, while the adversarial loss controlled the generator generating “real” images. Inspired by Context encoder, High-resolution inpainting method [46] inpainted high-resolution images by optimizing a joint loss combination including the holistic content constraint, the local texture constraint, and the TV loss. Pix2Pix [47] extended Context encoder into an end-to-end inpainting method with fine texture recovering, and Pix2PixHD [48] further enhanced Pix2Pix [47] in which the high-resolution street images were able to be inpainted with more fine details. Aiming to fulfill or change the textural details on a content image, neural style transferring [49–56] focused on transferring a formulated art style into a content image but maintaining the semantic content. Gatys et al. [49,50] started the neural style transferring by transferring the art style over the flow of ConvNet and formally defined the formulation of the art style offered by the style image – Gram matrix [49]. Johnson et al. [53] proposed a novel strategy to fast transfer style via an end-to-end ConvNet in which they formulated the perceptual loss of transferring textual details derived from the Gram matrix [49]. Chen et al. [55] inherited the formulated definition Gram matrix of textural details, and they further redefined the texton with the original style loss.

Image detail improvements: Aside from semantically inpainting to recover the semantic content, some researches focused on reconstructing the high-frequency details of a high-resolution image from its low-resolution counterpart, so the highly challenging task is converted into estimating the details [57]. SRCNN [58] offered a fundamental method to reconstruct high-resolution images

from their low-resolution counterparts by minimizing the L2 loss between them in a ConvNet. VDSR [59] learned the residual of the generated image minus the input low-resolution image and fulfilled the high-resolution image using the residuals. SRGAN [57] employed multiple residual blocks to learn deep features and combined the reconstruction loss and the adversarial loss to recover the high-resolution quality of generated images. High-resolution reconstruction using ConvNet exploits novel strategies to generate high-quality textural details while maintaining the original semantic content, which sufficiently enhances image details and inspires image generation tasks.

3. Fully-convolutional image inpainting

In this section, we introduce the pipeline of our model, including framework design, joint loss function, and some implementation details. In Section 3.1, we briefly give the holistic overview of our framework, and in Section 3.2, the design of the content network is illustrated. Joint loss function is shown in Section 3.3 and we eventually give some implementation details in Section 3.4.

3.1. Framework overview

First, we start our introduction with defining the representations of the variables and formulating a general framework overview: Given a corrupt image x, the inpainting model intends to infer the missing regions in which the semantic low-frequency silhouette and high-frequency details are reasonably restored. Here x is referred to as the corrupt image (or context) and $\hat{y}$ is the restored counterpart of x; y, by the way, is the corresponding ground truth of x. In our case, the corrupt context x can also be interpreted by a real-valued tensor x of size $w \times h \times c$ that denotes the context width w, the context height h, and the color channel c; additionally, a tensor $\hat{y}$ with the same size denotes the output of the model, a tensor y with the same size denotes the ground truth, and $x^{(i)}$ denotes the i-th intermediate feature map, respectively.

We ultimately aim to customize an end-to-end image inpainting method f_θ that takes as input a corrupt image context x and outputs an estimation $\hat{y} = f_\theta(x)$ with respect to x. By minimizing the predefined loss $\mathcal{L}$ between the estimation $\hat{x}$ and its corresponding latent ground truth y, errors of the semantic and high-frequency details are calculated and optimized so that f_θ can be obtained to inpaint corrupt context semantically. f_θ is trained jointly by the perceptual and the adversarial training, which can optimize the semantic error and reality of results respectively. The perceptual training can be formulated as a discrepancy between the output and its ground truth. In the adversarial training, the generator G_{θ_G} parametrized by θ_G and the discriminator D_{θ_D} parametrized by θ_D are customized to generate $\hat{y}$ that is indistinguishable for D_{θ_D}. For the training set $\{x_n; y_n\}_{n=1}^N$ the generator G_{θ_G} and discriminator D_{θ_D} are obtained by solving Eqs. (1) and (2) in which the joint loss $\mathcal{L}_{jointG}$ of generator G and $\mathcal{L}_{jointD}$ of discriminator D will be presented in Section 3.3:

$$\hat{\theta_G} = \text{argmin}_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{jointG}(G_{\theta_G}(x_n), y_n) \tag{1}$$

$$\hat{\theta_D} = \text{argmin}_{\theta_D} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{jointD}(D_{\theta_D}(x_n), y_n) \tag{2}$$

3.2. The content network

Encoder: Our encoder–decoder architecture is presented here. The multi-layer hierarchical convolutional net is adopted to extract semantic and detailed features. With the stride of 2, the input contexts are downsampled through all convolutional layers, and meanwhile, they are transformed into latent representations describing semantic features. To assist in accelerating the convergence when training, batch normalization layers follow corresponding convolutional layers one by one. The encoder is conventionally designed on account of the proved effectiveness.

Decoder: By using transpose convolution, upsampling the intermediate latent representations or feature maps is conventional in the traditional generator designing, whereas the checkerboard artifacts cannot be ignorable; it is attributed to the overlap in transposed convolution [60]. Alternatively, we alter the transpose convolutional layers with sub-pixel convolution layers [61] to upsample the latent representation of the corrupt context. Instead of upsampling latent feature representations by transpose convolution method with the fractional stride, the sub-pixel convolution is employed to fill the missing pixels by reshuffling activations within feature maps. The intermediate feature maps are calculated with below equation,

$$x^{(i+1)} = f^L(x^{(i)}) = \mathcal{PS}(W_L * x^{(i)} + b_L) \tag{3}$$

where f^L denotes the sub-pixel convolutional operation, W_L, b_L, and $*$ denote the convolutional kernel, the convolutional bias, and the convolutional operator in the sub-pixel convolutional operation f^L, and $\mathcal{PS}$ is the periodic channel-shuffling operator that rearranges channels of the previous feature maps $x^{(i)}$ and outputs upsampling feature maps $x^{(i+1)}$. The operator $\mathcal{PS}$ can be interpreted mathematically by below equation [61].

$$\mathcal{PS}(T)_{x,y,c} = T_{\lfloor x/r \rfloor, \lfloor y/r \rfloor, c \cdot r \cdot mod(y,r) + c \cdot mod(x,r)} \tag{4}$$

The results and comparisons will be evaluated and conducted in Section 4, where we will thoroughly explore the differences between our methods and the baseline.

Residual skip connections: For the ease of convergence when training and reuse of existing feature extracted in the encoder, residual skip connections are bridged over the encoder and the decoder. He et al. [7] has demonstrated that skip connections bridge the backpropagation of gradients across convolutional layers, and therefore, they can significantly extend the depth of the network. Additionally, skip connections enlarge the receptive fields to enhance the representation of the network. On the other hand, it can also mathematically account for the faster convergence of the net. The formulation of the residual skip connections are described in below equation.

$$x^{(i+1)} = \sigma\left(G_{\theta_G}(x^{(i)}) + x^{(N+1-i)}\right) \tag{5}$$

where x^i denotes the feature maps in layer i and σ is referred to as the nonlinear activation ReLU [3].

In addition to bridging and reusing the features propagated forward through the encoder and the decoder, residual skip connections propagate the gradients through the identity maps to accelerate convergence; hence, gradients can bypass intermediate non-linear activations to avoid the gradient vanishing. According to He et al. [7], ResNet achieved great success in image classification tasks of the ImageNet contest based on residual blocks, and the authors attributed the success to the extraction of latent features. Aiming to accelerate the convergence and stabilize training, we bridge residual skip connections in our generator, which is also illustrated in Fig. 2.

Multi-layer convolutional net: The multi-layer convolutional net consists of multiple residual blocks to sufficiently extract deep features, and the whole ConvNet promotes the high-frequency details of generated images. It has been demonstrated that Ledig et al. [57] offered a reference to reconstruct high-frequency details from its low-frequency counterpart, and we now confirm that the multi-layer ConvNet architecture is capable of smoothing the discontinuous boundaries. Therefore, the customized multi-layer ConvNet is

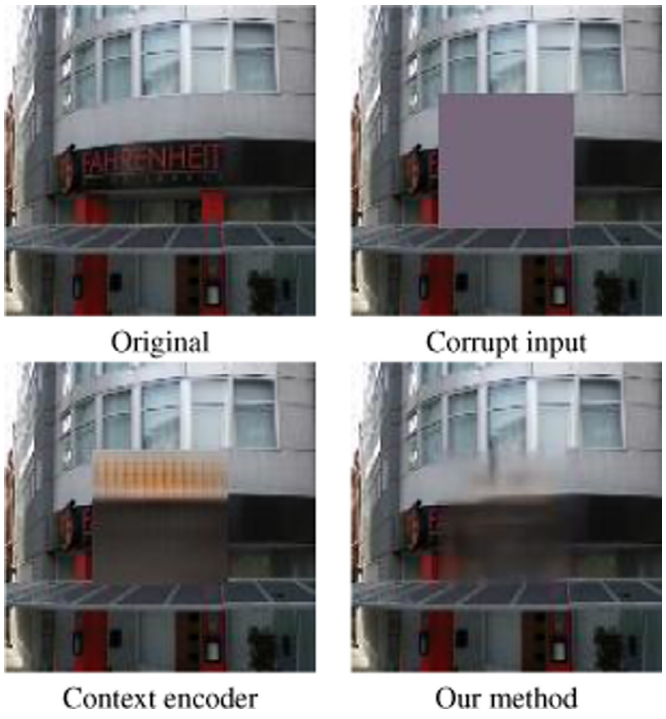

Fig. 1. Illustration of our recovery task. Concurrent inpainting models aim to predict a semantic filling region given a corrupt image. Our model can fix the corrupt hole with fine high-frequency details semantically and smoothly surrounding boundaries, whereas Context encoder [1] fails in recovering; the results of the Context encoder implementation by Kim [2] also illustrate the hallucination issues of retraining Context encoder.

employed to blur the discontinuous boundaries between the estimation and the context, on the condition of maintaining the high-frequency details. We will present the effectiveness of smoothing in the experiments.

Discriminator: Our ultimate goal of inpainting is to train a generator that can fill the missing regions with a semantically correct estimation and merge them seamlessly. Sole mean square error (MSE) as the reconstruction loss is widely used to measure the difference of the semantic content; the generated images, however, lack high-frequency details owing to MSE penalizing the distortion of the semantic content harder. A discriminator here is customized to penalize the generator generating low-resolution images, and gradients from the error offered by the discriminator can assist generator to adjust parameters. When successfully confusing the discriminator, the generator can produce high-quality images that have the same resolution as ones in the training set and its corresponding context. We conventionally denote the error from the discriminator as the adversarial loss and evaluate the effectiveness of the adversarial loss in the experiment.

3.3. Joint loss function

We conventionally use the combination of reconstruction loss (L2 loss) and perceptual loss (adversarial loss) to calibrate the prediction of our model. Here we describe the combination in this section.

We first define our holistic loss function that assists our model to predict missing semantic regions. The L2 loss function is conventionally employed because it can drive the net to reach the average of semantic errors, and thus the net can generate plausible images with the high-quality semantic silhouette. In Eq. (6), we define the L2 loss as the reconstruction loss.

$$\mathcal{L}_{L2}(G) = \mathbb{E}_{x,y}[||y - G(x)||_2^2] \tag{6}$$

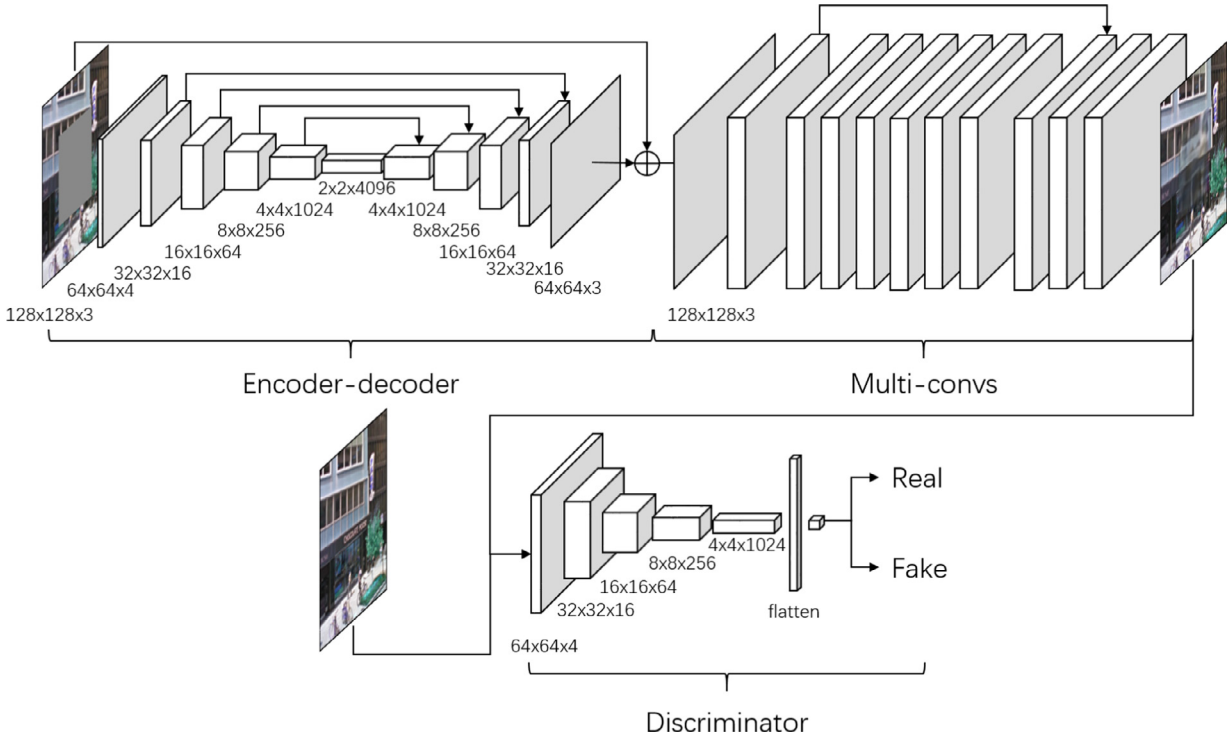

Fig. 2. The architecture of our inpainting network. We customize a generator composed of an encoder–decoder marked by "Encoder–decoder" and a multi-layer fully-convolutional net marked by "Multi-convs", which involves skip connections between intermediate feature maps and residual blocks. The discriminator marked by "Discriminator" is also illustrated in the figure. The lines illustrate skip connections between layers. The generator reuses front feature maps and smooths the boundaries and the discriminator offers the adversarial loss by differentiating if images come from the generator.

where y denotes the expected output and $G(x)$ denotes the output from the generator.

The L2 loss intends to recover the lost low-frequency holistic structure of the image, but if L2 loss dominates the optimization of training, it lacks high-frequency details so that the texture is inevitably blurred. Consequently, L2 loss function leads to overly smooth images with poor perceptual quality [53,62–64]. For enhancing the quality of high-frequency details, the adversarial loss is appended to our objective function. The adversarial loss pushes the net to generate images produced from the real image manifold [57]. Eq. (7) defines the GAN loss [32] as the objective function and the generator and the discriminator are obtained by solving the min-max game $\min_G \max_D \mathcal{L}_{GAN}(G, D)$.

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_y[\log D(y)] + \mathbb{E}_x[\log(1 - D(G(x)))] \tag{7}$$

where $D(\cdot)$, $G(\cdot)$, x, and y denote the discriminator, the generator, the corrupt context, and its corresponding expected output, respectively.

For the ease of implementation, the cost function of the generator are transformed to Eq. (8) by the "$-\log D$" trick [32] while the cost function of the discriminator remains solving Eq. (9).

$$\mathcal{L}_G(G, D) = \mathbb{E}_x[-\log D(G(x))] \tag{8}$$

$$\mathcal{L}_D(G, D) = -\mathbb{E}_y[\log D(y)] - \mathbb{E}_x[\log(1 - D(G(x)))] \tag{9}$$

By optimizing Eqs. (10) and (11) the optimal generator $\hat{G}$ and discriminator $\hat{D}$ can be numerically obtained.

$$\hat{G} = \text{argmin}_G \lambda_{L2}\mathcal{L}_{L2}(G) + \lambda_G\mathcal{L}_G(G, D) \tag{10}$$

$$\hat{D} = \text{argmin}_D \lambda_D\mathcal{L}_D(G, D) \tag{11}$$

where λ_{L2}, λ_G, and λ_D denote weights of components of the loss function to balance the adversarial and semantic learning.

Eventually, the objective function is attached with the reconstruction loss and the regularization term $\mathbb{E}[||W_G||_2]$ and $\mathbb{E}[||W_D||_2]$. Therefore the estimation of the generator $\hat{G}$ and the discriminator $\hat{D}$ can be obtained by minimizing the joint loss $\mathcal{L}_{jointG}$ of generator G and $\mathcal{L}_{jointD}$ of discriminator D.

$$\begin{aligned}\hat{G} &= argmin_G\mathcal{L}_{jointG}(G, D)\\ &= \text{argmin}_G \lambda_{L2}\mathcal{L}_{L2}(G) + \lambda_G\mathcal{L}_G(G, D) + \lambda_{regG}\mathbb{E}[||W_G||_2]\end{aligned} \tag{12}$$

$$\begin{aligned}\hat{D} &= argmin_D\mathcal{L}_{jointD}(G, D)\\ &= \text{argmin}_D \lambda_D\mathcal{L}_D(G, D) + \lambda_{regD}\mathbb{E}[||W_D||_2]\end{aligned} \tag{13}$$

where λ_{regG} and λ_{regD} denote weights of regularization terms balancing the penalty of the complexity of the model.

The pseudocode of training our recovery model is presented in Algorithm 1.

3.4. Implementation details

The default configurations of hyperparameters are set as follows. The default size of corrupt contexts is 128×128 or 256×256 and the layouts of the generator G and the discriminator D are already presented. The generator G and the discriminator D are optimized alternately and trained by the mini-batch ADAM [65] solver in which the learning rate is 0.002 and the momentum parameters β_1 is 0.5 and β_2 is 0.999. Default λ_G is 0.001 and the tradeoff of balancing the reconstruction loss and adversarial loss is thoroughly explored in the experiment. All the models tested in experiments are trained or retrained through over 100 epochs so that each model can be thoroughly optimized. At the inference, the generator G is executed independently to inpaint the corrupt context.

Algorithm 1 Minibatch ADAM training of ResContextEncoder for image inpainting.

Require: ~~

α: initial learning rate. m: batch size. n_{giter}: the number of iterations of training the generator. w_0: initial discriminator parameters. θ_0: initial generator parameters. λ_{L2}: reconstruction weight. λ_G: adversarial loss of generator. λ_{regG}: regularization weight of generator. λ_{regD}: regularization weight of discriminator.

while θ has not converge **do**
  Sample a batch $\{y^{(i)}\}_{i=1}^m \sim \mathbb{P}_r$ from training set.
  $\{x^{(i)}\}_{i=1}^m$, $\{x_{gt}^{(i)}\}_{i=1}^m \leftarrow \text{Crop}(\{y^{(i)}\}_{i=1}^m)$
  for $t = 1, \ldots, n_{giter}$ **do**
    $g_\theta \leftarrow \nabla_\theta[\lambda_{L2} \cdot \frac{1}{m}\sum_{i=1}^m (y^{(i)} - f_\theta(x^{(i)}))^2 + \lambda_G \cdot \frac{1}{m}\sum_{i=1}^m - \log f_w(f_\theta(x^{(i)})) + \lambda_{regG} \cdot \frac{1}{|\theta|}\sum_{\theta^{(i)} \in \theta}(\theta^{(i)})^2]$
    $\theta_t \leftarrow \text{ADAM}(\theta_{t-1}, g_\theta, \alpha)$
  end for
  $g_w \leftarrow \nabla_w[\lambda_D \cdot \frac{1}{m}\sum_{i=1}^m (-\log f_w(y^{(i)}) - \log(1 - f_w(f_\theta(x^{(i)})))) + \lambda_{regD} \cdot \frac{1}{|w|}\sum_{w^{(i)} \in w}(w^{(i)})^2]$
  $w \leftarrow \text{ADAM}(w, g_w, \alpha)$
end while

The generator can take as input a single context or a mini-batch of contexts and the same mini-batch is taken for the convenience of inference.

4. Experiments and results

To evaluate our model, we present experiments and results in this section. In Section 4.1, we exhibit some examples and quantitative results to illustrate and demonstrate the performance and the superiority of our method over the chosen baseline methods. In Section 4.2, we also exhibit some representative examples to illustrate the ablation study of components in our model. In Section 4.3, we aim to explore the effect of the ratio between reconstruction and adversarial loss, and meanwhile, we test the robustness of our model on this ratio. In Section 4.4, we evaluate the positive effect of estimating the missing regions in high-resolution images.

Datasets and quantitative indicators: The Street View Text (SVT) dataset [66,67] is adopted as our training and test set. The dataset contains 350 street view pictures depicting the buildings and billboards on the streets. A total of 300 pictures are preserved as the training set while the left pictures are preserved as the test set. The learning and generalization ability of models can be thoroughly evaluated since SVT is a small dataset. Quantitative results on the test set are reported and compared in terms of the mean square error (MSE), the peak signal to noise (PSNR) [68], and the structural similarity (SSIM) [69]. MSE is an essential indicator to evaluate the holistic discrepancy between our estimations and corresponding ground truths, while PSNR conventionally reflects the high-resolution details although it lacks semantic information. Additionally, SSIM assesses the perceptual quality of inpainting results. The quantitative indicators are computed in equations below.

$$MSE(y, \hat{y}) = \frac{1}{mnp}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\sum_{k=0}^{p-1}[y(i, j, k) - \hat{y}(i, j, k)]^2 \tag{14}$$

$$PSNR(y, \hat{y}) = 10 \times \log_{10}\frac{255^2}{MSE(y, \hat{y})} \tag{15}$$

$$SSIM(y,\hat{y}) = \frac{1}{p}\sum_{k=0}^{p-1}\frac{(2\mu_{y_k}\mu_{\hat{y}_k}+c_1)(2\sigma_{y_k\hat{y}_k}+c_2)}{(\mu_{y_k}^2+\mu_{\hat{y}_k}^2+c_1)(\sigma_{y_k}^2+\sigma_{\hat{y}_k}^2+c_2)} \tag{16}$$

where $\hat{y}$, y, m, n, and p denote the result, the ground truth, the width, the height, and the color channel of the result; μ, σ, c_1, and c_2 are the average, the variance, $(0.01\times 255)^2$ and $(0.03\times 255)^2$ by default. The paired t-test [70] is used to compare the significance of methods.

Baseline methods: Context encoder [1] and Pix2Pix [47] are reimplemented and retrained on the dataset as the baseline models, which are presented in CVPR 2016 and 2017, respectively, and the comparisons with Context encoder [1] and Pix2Pix [47] are conducted just because we aim to compare our method only with end-to-end methods, not with an optimal algorithm like High-resolution image inpainting [46].

4.1. Human perceptual and quantitative comparisons

Here we visually present the qualitative and quantitative comparison between the baseline methods and ours. We choose to reimplement and retrain Context encoder [1] and Pix2Pix [47] as the baseline methods. We first exhibit the visual performance from the baseline methods and ours to consolidate the superiority of our method: on the same training set, our method is capable of stabilizing training to convergence, but the baseline methods hardly reach the ideal convergence of the joint inpainting loss. Then we demonstrate the capability of generating reconstruction information and high-frequency details in terms of MSE, PSNR, and SSIM indicators. We persistently choose these three indicators to evaluate semantic reconstruction, high-frequency details, and structural similarity of generated image estimations since they are still well-accepted indicators for evaluation, even though they cannot reflect comprehensive semantic and high-frequency information.

We first present the qualitative results on the test set from the baseline methods and ours in Fig. 3. In Fig. 3, we can easily find out that our method can reconstruct more semantic information, yet the baseline methods suffer from the severe distortion of entities or tiling artifacts. We attribute the qualitative performance and superiority to the ability of residual learning: it should be of help when reconstructing images. We are not denying Context encoder [1] and Pix2Pix [47] can predict semantically correct and real images; however, it is notable that the baseline methods are not adequate to retrain to converge according to our experiments of implementing Context encoder [1] and Pix2Pix [47], and Yang et al. [46] has published some other evidence to prove that the inpainting results of Context encoder hardly generate fine textural details. From the qualitative results published in Fig. 3, we can also find out that Context encoder cannot generate proper textual details although it grabs the holistic image information. Results of our method are encouraging since they are semantically correct with sufficient textual details. In addition, we also find out in Fig. 4 that our method performs well on smoothing the discontinuous gaps between original contexts and generated missing regions, which should be attributed to the following multi-layer convolutional synthesis net in charge of merging them. Consequently, we can conclude that our method can predict semantically correct regions and synthesize those regions and original context seamless.

Aside from the distortion of semantic learning, the checkerboard artifacts are also presented in Fig. 3. Caused by the transposed convolution, checkerboard artifacts make the negative effort on the quality of generated images [60]. Context encoder [1] and Pix2Pix [47] both fail to entirely eliminate the artifacts, even if it can estimate the latent missing region. We replace transposed convolutions with the sub-pixel convolution so that our model can alleviate the checkerboard artifacts illustrated in Fig. 3. Thus, our method reconstructs images with better quality according to the test results.

Table 1
PSNR, MSE, and SSIM comparisons of the baseline methods, individual components and our proposed method ResCE.

Methods	Avg PSNR (db)	Avg MSE	Avg SSIM
CE [1] impl	23.12	370.24	0.8289
Pix2Pix [47]	19.37	825.73	0.7613
Multi-convs	22.29	452.05	0.8408
Subpx upsampling	20.88	573.64	0.8120
ResCE	**23.78**	**346.62**	**0.8532**

Averages of the quantitative indicators are listed in Table 1 to demonstrate the superiority of our method compared to the baseline methods quantitatively. MSE, PSNR, and SSIM are adopted to evaluate these results quantitatively: MSE is used to evaluate the holistic loss, PSNR indicates the quality of high-frequency details, and SSIM assesses the structural similarity. Table 1 shows that the averages of MSE, PSNR, and SSIM of our method are over those of Context encoder [1] and Pix2Pix [47] as the baseline methods, and significant differences between our method and the baseline methods are observed on both indicators since the p-values are less than 0.01 with the paired t-test. It can be seen that our method generally outperforms the baseline methods on our test dataset and the improvement is significant, which demonstrates the consistency of the superiority of our method.

Some exemplars are illustrated in Fig. 4 in which we zoom in the left upper corners of images and observe the edges more clearly. CE [1] and Pix2Pix [47] fails to smooth the edges, which leads to a degradation of merging images. Our method, however, can upgrade the quality by smoothing the edges. PSNR indicators of our method also exceed those of CE [1] and Pix2Pix [47], on the basis of Fig. 4. Since we evaluate and illustrate the holistic error and high-frequency details with MSE, PSNR, and SSIM in Table 1 and Fig. 4, we can conclude that our method can generate images with lower holistic errors and more high-frequency details.

4.2. Ablation study of components in the generator

We further analyze the powerful effect of each individual component of our method. Composed of the residual encoder–decoder and the multi-layer convolution net, our model ought to succeed in translating the latent representation into a reliable estimation of missing regions and modifying the discontinuous edges, so here we validate the effect of an individual component by eliminating the other one. Figs. 5 and 6, and Table 1 show the results of each individual component of our method qualitatively and quantitatively.

Fig. 5 intuitively shows the ablation study of the individual component of our model. The multi-layer convolutional net can fill the missing region with its surrounding pixels but is hard to infer the center semantically; it, however, is adequate to infer the surrounding continuous edges between the estimation of the missing region and original context. The sub-pixel residual encoder–decoder can capture the semantically correct structure and generate parts of high-frequency details, yet the discontinuous edges between the generated image and the original context distract the consistency of the visual quality. Our method takes advantage of these two components thoroughly and can synthesize these two parts into a semantically correct estimation of the missing region with smooth edges, which can fill the missing center seamlessly.

In Fig. 6, we zoom in the edges between the original contexts and generated missing regions, and can also find that combination can take advantages of these two components to fill the missing regions and smooth the discontinuous edges semantically. Note that the combination achieves the highest PSNR, MSE, and SSIM

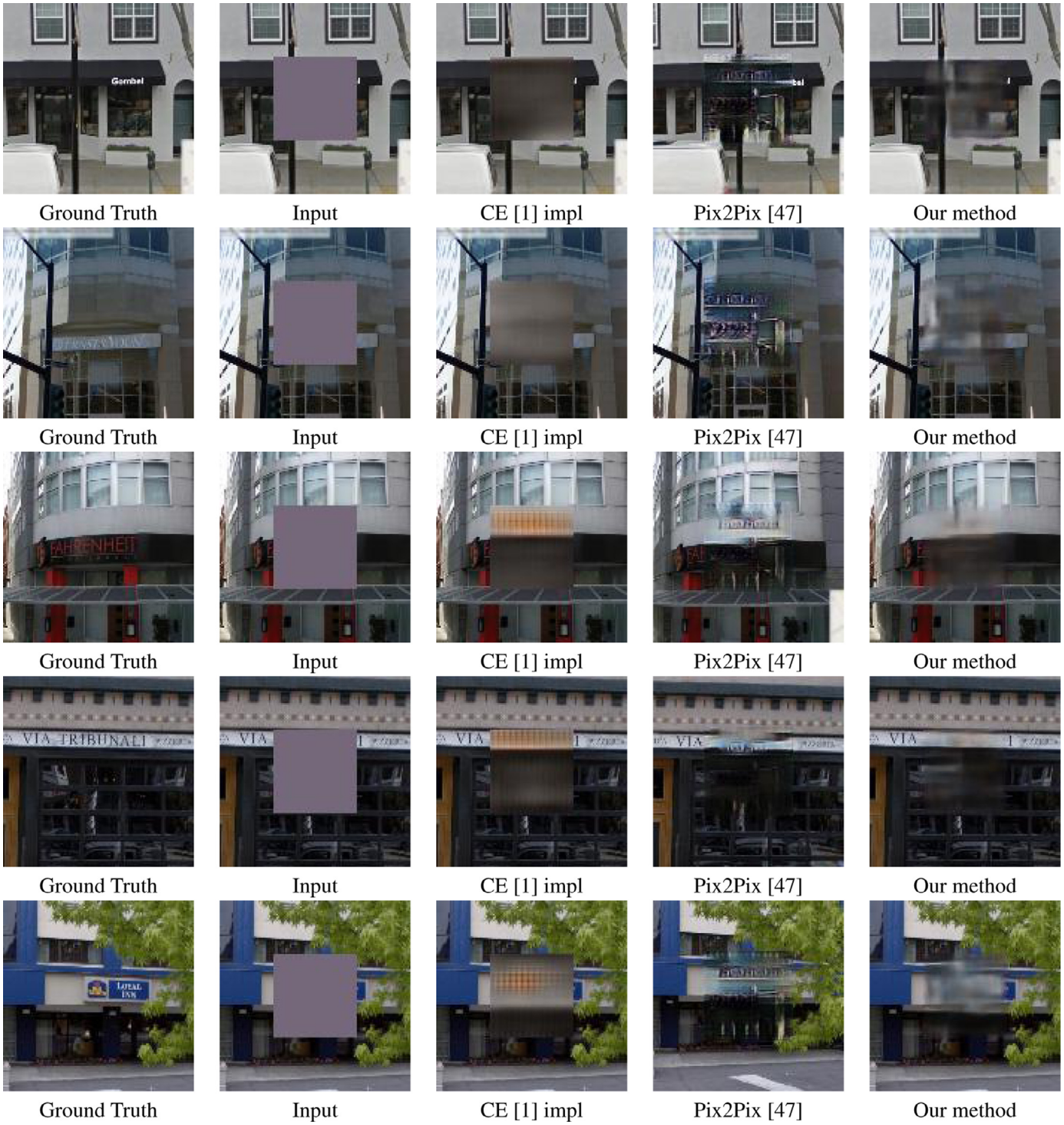

Fig. 3. Comparisons between the baseline methods and ours. Examples are exhibited here to intuitively evaluate the effectiveness of semantic inpainting of the baseline methods and ours. From these examples, the superiority of our method over the baseline can be easily found out.

according to Table 1 and Fig. 6, which can demonstrate the combination generally outperforms each individual on the test dataset since significant differences are observed with the p-values < 0.01 with the paired t-test.

4.3. Robustness of the weight tradeoff between adversarial loss and reconstruction loss

In this section, we test the robustness of the weight tradeoff between adversarial loss and reconstruction loss. The adversarial loss can enhance the realism of the generated images, and the reconstruction loss tends to penalize the semantic discrepancy between the inference and its corresponding ground truth. Higher adversarial loss, however, can lead to higher semantic errors, while lower adversarial loss disables generating high-resolution images. We, therefore, test if our method is robust on the weight tradeoff between these losses, and Fig. 7 and Table 2 illustrate the results and conclusions regarding the tradeoff.

In Table 2, the averages of PSNR with respect to $\lambda_{adv} = 0.001, 0.01, 0.05$ are 23.78 db, 23.85 db, and 23.61 db, but no

Fig. 4. Comparisons of PSNR between the baseline methods and ours. We zoom in the left upper corners and observe the edges over there. CE [1] and Pix2Pix [47] fail to smooth the edges, which leads to discontinuous boundaries that decrease the perception of human beings. Whereas, our method upgrades smoothing the edges. PSNR is commonly used to indicate the high-frequency performance of inpainting methods, and we can find out that our method exceeds CE [1] and Pix2Pix [47]. Consequently, it is convincing that our method outperforms CE [1] and Pix2Pix [47] qualitatively and quantitatively in terms of PSNR.

Table 2
The tradeoff of weight of adversarial loss and the quantitative indicators with respect to HR (256×256) inpainting.

Methods	Avg PSNR (db)	Avg MSE	Avg SSIM
Ours / $\lambda_{adv} = 0.001$	**23.78**	346.62	0.8532
Ours / $\lambda_{adv} = 0.01$	**23.85**	335.12	0.8517
Ours / $\lambda_{adv} = 0.05$	**23.61**	353.15	0.8475
Ours / 256×256	23.06	402.10	0.8506

significant differences are observed to support the solid improvement of models with a greater weight of adversarial loss (p-value $\geq$ 0.05). Some examples are also presented in Fig. 7, and the holistic inpainting results are coherent as PSNR shows even if there are mere visual differences between them. So we can conclude that a greater weight of the adversarial loss makes less effect on the results in terms of illustrations and PSNR. We further conclude that our method can be robust on the weight of adversarial loss,

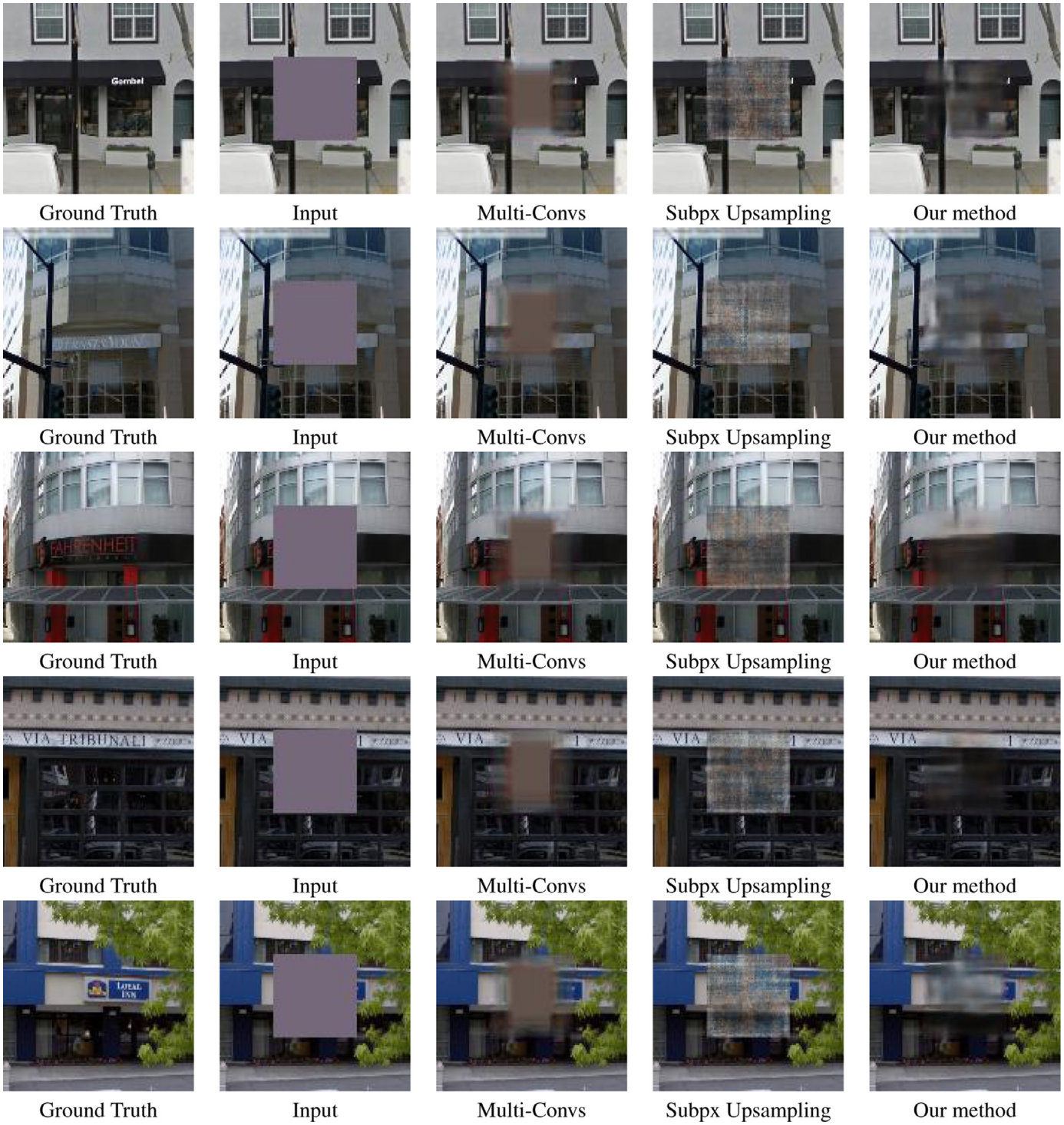

Fig. 5. Examples of results from individual components of our method. Each individual component is tested to generate the missing center regions, and the results are presented here. Our encoder–decoder architecture can semantically generate the missing areas, but the discontinuous gap between original and generated images causes visual disharmony. Our multi-layer convolution architecture cannot infer semantically missing regions but is capable of smoothing the boundaries between original and generated images. Naturally, the combination may inherit these benefits from each individual component, which is demonstrated and illustrated by the results in the last column.

i.e., our model can generate stable inpainting results given whatever weight of adversarial loss.

4.4. Multi-scale image inpainting

We study the generalization of our model on multi-scale image inpainting tasks, and we test our model on the Street View Text dataset rescaled to 256×256. The quantitative results are listed in Table 2, the inpainting results are exhibited in Fig. 8, and PSNR indicators are listed below.

We can still find out that our model can recover missing regions with respect to the corrupt context semantically and correctly. Our model can first predict missing horizontal and vertical lines to fabricate the components and modify them with

Fig. 6. Comparisons of PSNR between individual components and the proposed method. PSNR is used to evaluate each individual component quantitatively, and the combination achieves the best inpainting quality in terms of PSNR.

the surrounding texture, which recovers the reality of the corrupt image. The multi-layer synthesis convolutional net smooths the discontinuous gaps between the original context and its corresponding predicted region, and consequently, these two parts are synthesized together seamless. The MSE, PSNR, and SSIM indicators in Table 2 also demonstrate that our model can recover them in high-resolution quality. These high-quality inpainting results are attributed to the nature of our method – our method outperforms on extracting semantic features and inferring semantically correct missing regions compared to Context encoder [1] and Pix2Pix [47] based on residual skip connections; multi-layer synthesis convolutional net also helps to smooth the discontinuous gap between the original context and its prediction.

5. Discussions and conclusions

In this paper, we investigate a multi-layer convolutional model for inpainting the missing region semantically and synthesizing inference seamlessly in a corrupt context, and we demonstrate that skip connections built between layers are capable of improving

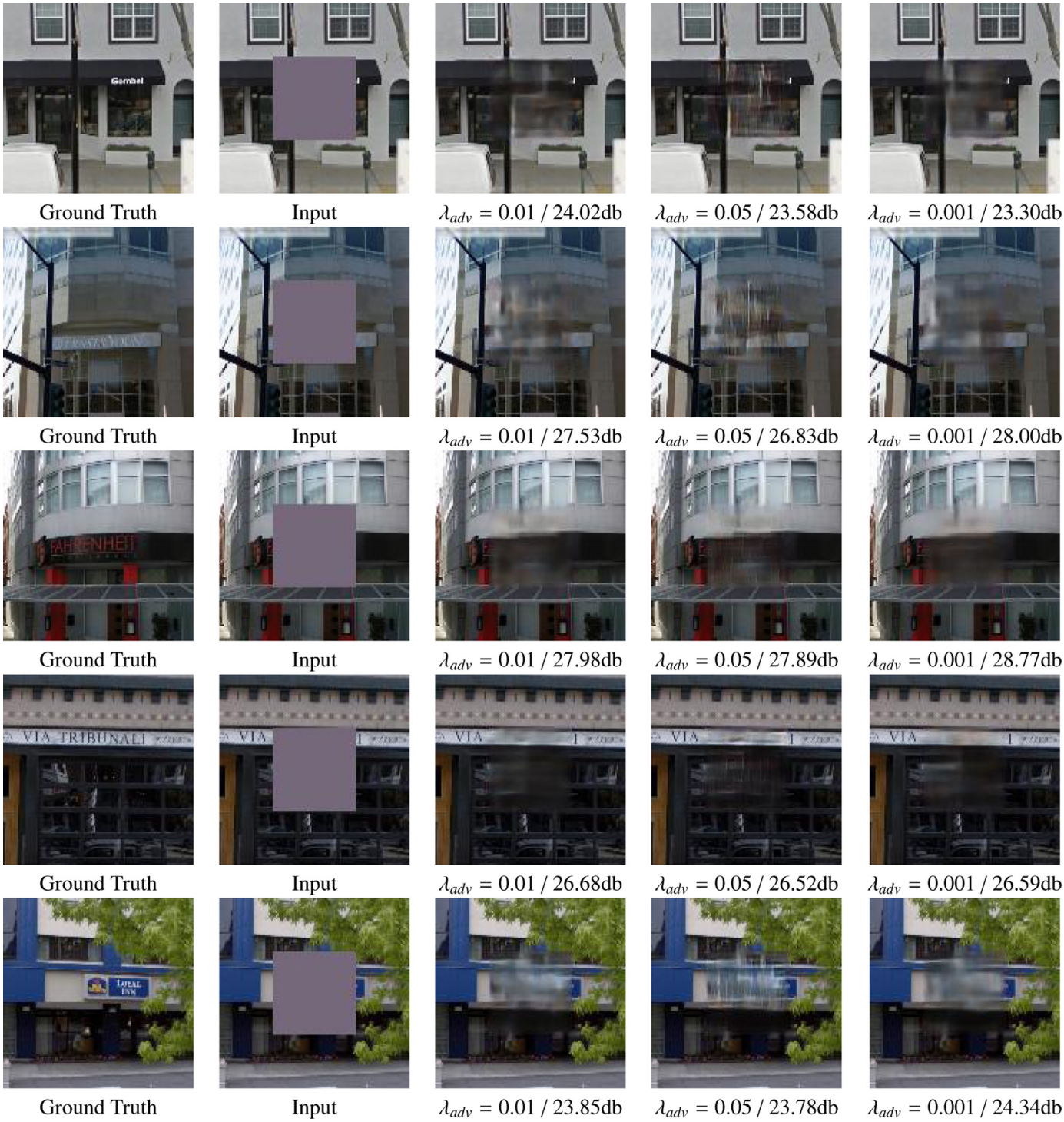

Fig. 7. Comparison of weights of adversarial loss. Slightly, larger weight leads to sharper results while smaller weight results in more blur results. However, all of our methods with different weights achieve acceptable inpainting results even if there are mere visual differences between them (p-value $\geq$ 0.01).

semantic inference while the multi-convolutional net can smoothing the discontinuous edges. To achieve this, we propose a feasible prototype with skip connections and a multi-layer convolutional smoothing component and test its performance on a real-world building dataset. According to the experiments and results, conclusions can be summarized as follows:

1. Our method outperforms the baseline methods in predicting the missing region semantically and synthesizing seamlessly; it can improve the resolution quality of inpainting counterparts effectively as well. The quantitative results have convincingly demonstrated the superiority and significance in terms of MSE, PSNR, and SSIM.
2. The ablation study illustrates and indicates the functionality of each individual component in our model: the encoder–decoder with residual learning and subpixel upsampling infers semantically and the multi-convolutional net smooths the boundaries. The combination of two parts improves the inference significantly in terms of qualitative and quantitative comparisons.

Fig. 8. Inpainting Results of High-resolution Images (256 × 256). Reconstruction images are rescaled to fit the format. Ground truths, results of our method, and upscaling results are illustrated to demonstrate the inpainting performance. PSNR is used here to indicate the high-frequency performance. Our method can achieve the acceptable performance in terms of human perception and the quantitative indicator.

3. The study concerning the weight of adversarial loss illustrates and indicates that our method can be robustly trained and be more tolerant of the ratio of reconstruction and adversarial loss.
4. With minor fine-tuning, our model can be adopted in inpainting images with higher-resolution, which demonstrates the extension and generalization of our model.

In future, we shall further investigate extending our model to more high-resolution image inpainting since we need to qualitatively and quantitatively demonstrate our model's learning and inferring ability.

Acknowledgments

This research is partially sponsored by National Natural Science Foundation of China (Nos. 61571049, 61472044, 61472403, 61601033), the BNU Graduate Students' Platform for Innovation & Entrepreneurship Training Program (No. 3122121F1), the Fundamental Research Funds for the Central Universities (2016NT14), and by SRF for ROCS, SEM.

References

[1] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: feature learning by inpainting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.
[2] T. Kim: Implementation of "Context Encoders: Feature Learning by Inpainting", 2016, https://github.com/jazzsaxmafia/Inpainting.
[3] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[4] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Proceedings of European Conference on Computer Vision, Springer, 2014, pp. 818–833.
[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, et al., Going deeper with convolutions, in: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[6] K. Simonyan, A. Zisserman, Very dezep convoluztional Networks for Large-scale imzage Recognition, (2014) arXiv:1409.1556.
[7] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[8] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 5987–5995.
[9] G. Larsson, M. Maire, G. Shakhnarovich, Fractalnet: Ultra-deep Neural Networks Without Residuals, (2016) arXiv:1605.07648.
[10] G. Huang, Z. Liu, K.Q. Weinberger, L. van der Maaten, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1, 2017, p. 3.
[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFS, (2016) arXiv:1606.00915.
[12] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[13] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1520–1528.
[14] E.L. Denton, S. Chintala, R. Fergus, et al., Deep generative image models using

a laplacian pyramid of adversarial networks, in: Proceedings of Advances in Neural Information Processing Systems, 2015, pp. 1486–1494.
[15] L. Theis, M. Bethge, Generative image modeling using spatial lstms, in: Proceedings of Advances in Neural Information Processing Systems, 2015, pp. 1927–1935.
[16] K. Gregor, I. Danihelka, A. Graves, D.J. Rezende, D. Wierstra, Draw: A Recurrent Neural Network for Image Generation, (2015) arXiv:1502.04623.
[17] D.J. Im, C.D. Kim, H. Jiang, R. Memisevic, Generating Images with Recurrent Adversarial Networks, (2016) arXiv:1602.05110.
[18] A. Dosovitskiy, T. Brox, Inverting visual representations with convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4829–4837.
[19] A.v.d. Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel Recurrent Neural Networks, (2016) arXiv:1601.06759.
[20] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 2980–2988.
[21] Y.-W. Tai, S. Liu, M.S. Brown, S. Lin, Super resolution using edge prior and single image detail synthesis, in: Proceedings of 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 2400–2407.
[22] J. Sun, Z. Xu, H.-Y. Shum, Image super-resolution using gradient profile prior, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2008, pp. 1–8.
[23] K. Zhang, X. Gao, D. Tao, X. Li, Multi-scale dictionary for single image super-resolution, in: Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 1114–1121.
[24] H. Yue, X. Sun, J. Yang, F. Wu, Landmark image super-resolution by retrieving web images, IEEE Trans. Image Process. 22 (12) (2013) 4865–4878.
[25] R. Timofte, V. De, L. Van Gool, Anchored neighborhood regression for fast example-based super-resolution, in: Proceedings of 2013 IEEE International Conference on Computer Vision (ICCV), IEEE, 2013, pp. 1920–1927.
[26] R. Timofte, V. De Smet, L. Van Gool, A+: adjusted anchored neighborhood regression for fast super-resolution, in: Proceedings of Asian Conference on Computer Vision, Springer, 2014, pp. 111–126.
[27] K.I. Kim, Y. Kwon, Single-image super-resolution using sparse regression and natural image prior, IEEE Trans. Pattern Anal. Mach. Intell. 32 (6) (2010) 1127–1133.
[28] H. He, W.-C. Siu, Single image super-resolution using Gaussian process regression, in: Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 449–456.
[29] J. Salvador, E. Perez-Pellitero, Naive Bayes super-resolution forest, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 325–333.
[30] S. Schulter, C. Leistner, H. Bischof, Fast and accurate image upscaling with super-resolution forests, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3791–3799.
[31] D. Dai, R. Timofte, L. Van Gool, Jointly optimized regressors for image super-resolution, in: Proceedings of Computer Graphics Forum, 34, Wiley Online Library, 2015, pp. 95–104.
[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[33] M. Mirza, S. Osindero, Conditional Generative Adversarial Nets, (2014) arXiv:1411.1784.
[34] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, Infogan: interpretable representation learning by information maximizing generative adversarial nets, in: Proceedings of Advances in Neural Information Processing Systems, 2016, pp. 2172–2180.
[35] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired Image-to-image Translation Using Cycle-consistent Adversarial Networks, (2017) arXiv:1703.10593.
[36] Z. Yi, H. Zhang, P.T. Gong, et al., Dualgan: Unsupervised Dual Learning for Image-to-image Translation, (2017) arXiv:1704.02510.
[37] T. Kim, M. Cha, H. Kim, J. Lee, J. Kim, Learning to Discover Cross-domain Relations with Generative Adversarial Networks, (2017) arXiv:1703.05192.
[38] A. Radford, L. Metz, S. Chintala, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, (2015) arXiv:1511.06434.
[39] H. Zhang, Y. Sun, L. Liu, X. Wang, L. Li, W. Liu, Clothingout: a category-supervised gan model for clothing segmentation and retrieval, Neural Comput. Appl. (2018), doi:10.1007/s00521-018-3691-y.
[40] Y. Ji, H. Zhang, Q.J. Wu, Saliency detection via conditional adversarial image–to-image network, Neurocomputing 316 (2018) 357–368.
[41] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein Gan, (2017) arXiv:1701.07875.
[42] M. Arjovsky, L. Bottou, Towards Principled Methods for Training Generative Adversarial Networks, (2017) arXiv:1701.04862.
[43] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved Training of Wasserstein Gans, (2017) arXiv:1704.00028.
[44] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S.P. Smolley, Least squares generative adversarial networks, in: Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 2813–2821.
[45] G.-J. Qi, Loss-sensitive Generative Adversarial Networks on Lipschitz Densities, (2017) arXiv:1701.06264.
[46] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, H. Li, High-resolution Image Inpainting Using Multi-scale Neural Patch Synthesis, (2016) arXiv:1611.09969.
[47] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image Translation with Conditional Adversarial Networks, (2016) arXiv:1611.07004.
[48] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution Image Synthesis and Semantic Manipulation with Conditional gzans, (2017) arXiv:1711.11585.
[49] L.A. Gatys, A.S. Ecker, M. Bethge, A Neural Algorithm of Artistic Style, (2015a) arXiv:1508.06576.
[50] L.A. Gatys, A.S. Ecker, M. Bethge, Texture Synthesis and the Controlled genezration of natuzral Stimuli Using Convolutional Neural Networks, 12 (2015b) arXiv:1505.07376.
[51] A.J. Champandard, Semantic Style Transfer and Turning Two-bit Doodles into Fine Artworks, (2016) arXiv:1603.01768.
[52] D. Ulyanov, V. Lebedev, A. Vedaldi, V.S. Lempitsky, Texture networks: feed–forward synthesis of textures and stylized images., in: Proceedings of International Conference on Machine Learning (ICML), 2016, pp. 1349–1357.
[53] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: Proceedings of European Conference on Computer Vision, Springer, 2016, pp. 694–711.
[54] C. Li, M. Wand, Combining markov random fields and convolutional neural networks for image synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2479–2486.
[55] D. Chen, L. Yuan, J. Liao, N. Yu, G. Hua, Stylebank: An Explicit Representation for Neural Image szyle Transfer, (2017) arXiv:1703.09210.
[56] J. Liao, Y. Yao, L. Yuan, G. Hua, S.B. Kang, Visual Attribute Transfer through Deep Image Analogy, (2017) arXiv:1705.01088.
[57] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic Single Image Super-resolution Using a Generative Adversarial Network, (2016) arXiv:1609.04802.
[58] C. Dong, C.C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in: Proceedings of European Conference on Computer Vision, Springer, 2014, pp. 184–199.
[59] J. Kim, J. Kwon Lee, K. Mu Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1646–1654.
[60] A. Odena, V. Dumoulin, C. Olah, Deconvolution and checkerboard artifacts, Distill 1 (10) (2016) e3.
[61] W. Shi, J. Caballero, F. Huszár, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1874–1883.
[62] M. Mathieu, C. Couprie, Y. LeCun, Deep Multi-scale Video Prediction Beyond Mean Square Error, (2015) arXiv:1511.05440.
[63] A. Dosovitskiy, T. Brox, Generating images with perceptual similarity metrics based on deep networks, in: Proceedings of Advances in Neural Information Processing Systems, 2016, pp. 658–666.
[64] J. Bruna, P. Sprechmann, Y. LeCun, Super-resolution with Deep Convolutional Sufficient Statistics, (2015) arXiv:1511.05666.
[65] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, (2014) arXiv:1412.6980.
[66] K. Wang, B. Babenko, S. Belongie, End-to-end scene text recognition, in: Proceedings of 2011 IEEE International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 1457–1464.
[67] K. Wang, S. Belongie, Word spotting in the wild, in: Proceedings of European Conference on Computer Vision, Springer, 2010, pp. 591–604.
[68] A. Hore, D. Ziou, Image quality metrics: PSNR vs. SSIM, in: Proceedings of 2010 20th International Conference on Pattern Recognition (ICPR), IEEE, 2010, pp. 2366–2369.
[69] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[70] T.N. Raju, William sealy gosset and william a. silverman: two “students” of science, Pediatrics 116 (3) (2005) 732–735.

Libin Jiao received his B.E. degree in College of Information Science and Technology from Beijing Normal University, China in 2014. He is currently pursuing Ph.D. degree in College of Information Science and Technology, Beijing Normal University, China. His research interests include Data Mining, Machine Learning, and Computer Vision.

Hao Wu is currently an instructor in College of Information Science and Technology, Beijing Normal University. He received the B.E. and Ph.D. degrees from Beijing Jiaotong University, Beijing, China, in 2010 and 2015, respectively. From October 2013 to April 2015, he worked as a research associate in Lawrence Berkeley National Laboratory. Until now, he still takes charge of some related research projects in Lawrence Berkeley National Laboratory. His research interests include data mining, machine learning, and computer vision.

Haodi Wang is currently a post-graduate student in Beijing Normal University, where she got her Bachelor degree and is making effort to get her Master degree. She devoted herself to the research of deep learning and Inpainting.

Rongfang Bie is currently a Professor at the College of Information Science and Technology of Beijing Normal University where She received her M.S. degree on June 1993 and Ph.D. degree on June 1996. She was with the Computer Laboratory at the University of Cambridge as a visiting faculty from March 2003 for one year. She is the author or co-author of more than 100 papers. Her current research interests include deep learning, knowledge representation and acquisition for the Internet of Things, dynamic spectrum allocation, big data analysis and application etc.