

Transposer: Universal Texture Synthesis Using Feature Maps as Transposed Convolution Filter

GUILIN LIU, NVIDIA

ROHAN TAORI, NVIDIA, UC Berkeley

TING-CHUN WANG, NVIDIA

ZHIDING YU and SHIQIU LIU, NVIDIA

FITSUM A. REDA and KARAN SAPRA, NVIDIA

ANDREW TAO and BRYAN CATANZARO, NVIDIA



Fig. 1. The first row shows our 4×4 times larger texture synthesis results. The remaining two rows show the texture synthesis results using different approaches. Transposer (ours) represents our method, which is generalizable and can perform texture synthesis on unseen texture images with a single network forward pass in tens or hundreds of milliseconds. Self-Tuning [Kaspar et al. 2015] sometimes fails to fully preserve the regular structure and needs hundreds of seconds to solve the objective function. pix2pixHD [Wang et al. 2018] simply enlarges the input rather than perform synthesis. Texture Mixer [Yu et al. 2019] cannot handle inputs with structures. sinGAN [Shaham et al. 2019] and Non-stat. [Zhou et al. 2018] need to take tens of minutes or several hours retrain their models for each input texture. WCT [Li et al. 2017a], the style transfer based method can't preserve structure patterns. DeepTexture [Gatys et al. 2015a] is optimization based method and needs tens of minutes.

Conventional CNNs for texture synthesis consist of a sequence of (de-)convolution and up/down-sampling layers, where each layer operates locally and lacks the ability to capture the long-term structural dependency required by texture synthesis. Thus, they often simply enlarge the input texture, rather than perform reasonable synthesis. As a compromise, many recent methods sacrifice generalizability by training and testing on the same single (or fixed set of) texture image(s), resulting in huge re-training time

Authors' addresses: Guilin Liu, NVIDIA; Rohan Taori, NVIDIA, UC Berkeley; Ting-Chun Wang, NVIDIA; Zhidong Yu; Shiqiu Liu, NVIDIA; Fitsum A. Reda; Karan Sapra, NVIDIA; Andrew Tao; Bryan Catanzaro, NVIDIA.

costs for unseen images. In this work, based on the discovery that the assembling/stitching operation in traditional texture synthesis is analogous to a transposed convolution operation, we propose a novel way of using transposed convolution operation. Specifically, we directly treat *the whole encoded feature map of the input texture as transposed convolution filters* and *the features' self-similarity map*, which captures the auto-correlation information, as *input to the transposed convolution*. Such a design allows our framework, once trained, to be generalizable to perform synthesis of unseen textures with a single forward pass in nearly real-time. Our method achieves state-of-the-art texture synthesis quality based on various metrics. While self-similarity helps preserve the input textures' regular structural patterns, our framework can also take random noise maps for irregular input textures instead of self-similarity maps as transposed convolution inputs. It allows to

get more diverse results as well as generate arbitrarily large texture outputs by directly sampling large noise maps in a single pass as well.

CCS Concepts: • Computing methodologies → Texturing.

Additional Key Words and Phrases: texture Synthesis; transposed Convolution, generalizability

ACM Reference Format:

Guilin Liu, Rohan Taori, Ting-Chun Wang, Zhiding Yu, Shiqiu Liu, Fitzum A. Reda, Karan Sapra, Andrew Tao, and Bryan Catanzaro. 2020. Transposer: Universal Texture Synthesis Using Feature Maps as Transposed Convolution Filter. 1, 1 (July 2020), 16 pages. <https://doi.org/10.1145/nnnnnnnnnnnnnn>

1 INTRODUCTION

Texture synthesis is defined as the problem of generating a large image output given a small example input such that the visual features and structures are preserved both locally and globally. Many methods have been explored in the past two decades including pixel-based methods [Efros and Leung 1999], assembling based methods [Efros and Freeman 2001; Kwatra et al. 2003], optimization based methods [Kaspar et al. 2015; Kwatra et al. 2005], etc.

Inspired by the unprecedented success of deep learning in computer vision, others have explored deep learning methods for texture synthesis. Existing works fall into one of two categories. Either an optimization procedure is used to match deep feature statistics in a pre-trained network [Gatys et al. 2015a; Li et al. 2017a], resulting in a slow generation process; or a network is trained to overfit on a fixed image or set of images [Li et al. 2017b; Shaham et al. 2019; Zhou et al. 2018], which prevents it from generalizing to unseen textures and needs to spend huge re-training time for every unseen texture image.

One reason for the bad generalization ability of these aforementioned methods [Li et al. 2017b; Shaham et al. 2019; Zhou et al. 2018] is because these one-model-per-image (set) approaches usually employ conventional image-to-image translation networks, which first embed the input into a feature space and then fully rely on a sequence of upsampling and convolutional layers to reach the target output size. Each upsampling and convolutional layer is a local operation lacking of global awareness. This design works well for tasks such as image super resolution, where the task is to enhance or modify local details. However, texture synthesis differs from super resolution in that texture synthesis, when viewed from a classical perspective, involves displacing and assembling copies of the input texture using different optimal offsets in a seamless way. The optimal displacement and assembling strategy involves much longer-range operations and compatibility checking, which are usually not easy to model with the conventional design by fully relying on a sequence of local up/down-sampling and (de)convolutional layers.

In the column pix2pixHD of Figure 1, we show that a conventional image-to-image translation network adapted from pix2pixHD [Wang et al. 2018] fails to perform reasonable texture synthesis, but instead mostly just enlarges the local contents for the input textures even though it has been trained to convergence using the same input and output pairs as our method.

In this paper, we propose a new deep learning based texture synthesis framework that generalizes to arbitrary unseen textures and synthesizes larger-size outputs. From a classical view, the texture

synthesis task can also be interpreted as the problem of first finding an appropriate offset to place a copy of the input texture image and then using optimization technique to find the optimal seam between this newly placed copy and the existing image to assemble them together. Our method follows some similar spirits but have some major differences in the following manner: 1) We perform assembling in feature space and at multiple scales; 2) The optimal shifting offset and assembling weights are modeled with the help of a score map, which captures the similarity and correlation between different regions of the encoded texture image. We call this score map a *self-similarity map* (discussed in details in Section 3); 3) We later show that the shifting and assembling operations can be efficiently performed with a single forward pass of a *transposed convolution operation* [Dumoulin and Visin 2016], where we *directly use the encoded feature of input textures as transposed convolution filters, and the self-similarity map as transposed convolution input*. Unlike traditional transposed convolution, our transposed convolution filters are not learnable parameters. While self-similarity map plays a key role in preserving the regular structural patterns, alternately, our framework also allows to take random noise map as input instead of self-similarity map to generate diverse outputs and arbitrarily large texture output with a single shot by accordingly sampling large random noise map for irregular structural texture inputs.

In this work, we make the following contributions: 1) We present a generalizable texture synthesis framework that performs faithful synthesis on unseen texture images in nearly real time with a single forward pass. 2) We propose a self-similarity map that captures the similarity and correlation information between different regions of a given texture image. 3) We show that the shifting and assembling operations in traditional texture synthesis methods can be efficiently implemented using a transposed convolution operation. 4) We achieve state-of-the-art texture synthesis quality as measured by existing image metrics, metrics designed specifically for texture synthesis, and in user study. 5) We show that our framework is also able to generate diverse and arbitrarily large texture synthesis results by sampling random noise maps.

2 RELATED WORK

We provide a brief overview of the existing texture synthesis methods. A complete texture synthesis survey can be found in [Wei et al. 2009], which is out of the scope of this work.

Non-parametric Texture Synthesis. Existing texture synthesis methods include pixel-based methods [Efros and Leung 1999; Wei and Levoy 2000], assembling based methods [Efros and Freeman 2001; Kwatra et al. 2003; Liang et al. 2001; Pritch et al. [n.d.]], optimization based methods [Kaspar et al. 2015; Kwatra et al. 2005; Portilla and Simoncelli 2000; Rosenberger et al. 2009], appearance space synthesis [Lefebvre and Hoppe 2006], etc. There are also some other works [Hertzmann et al. 2001; Lefebvre and Hoppe 2006; Rosenberger et al. 2009; Wu and Yu 2004; Wu et al. 2013; Zhang et al. 2003] showing interesting synthesis results; however, they usually need additional user manual inputs.

Among these traditional methods, self-tuning texture optimization [Kaspar et al. 2015] is the current state-of-the-art method. It uses

image melding [Darabi et al. 2012] with automatically generated and weighted guidance channels, which helps to reconstruct the middle-scale structures in the input texture. Our method is motivated by assembling based methods. [Kwatra et al. 2003] is a representative method of this kind, where texture synthesis is formulated as a graph cut problem. The optimal offset for displacing the input patch and the optimal cut between the patches can be found by solving the graph cut objective function, which sometimes could be slow.

Deep Feature Matching-based Texture Synthesis. Traditional optimization based methods [Kaspar et al. 2015; Kwatra et al. 2005; Portilla and Simoncelli 2000; Rosenberger et al. 2009] rely on matching the global statistics of the hand-crafted features defined on the input and output textures. Recently, some deep neural networks based methods have been proposed as a way to use the features learned from natural image priors to guide the optimization procedure. Gatys et al. [Gatys et al. 2015a] define the optimization procedure as minimizing the difference in gram matrices of the deep features between the input and output texture images. Sendik et al. [Sendik and Cohen-Or 2017] and Liu et al. [Liu et al. 2016] modify the loss proposed in [Gatys et al. 2015a] by adding a structural energy term and a spectrum constraint, respectively, to generate structured and regular textures. However, in all cases, these optimization-based methods are prohibitively slow due to the iterative optimizations.

Learning-based Texture Synthesis. Johnson et al. [Johnson et al. 2016] and Ulyanov [Ulyanov et al. 2016] alleviate the previously mentioned optimization problem by training a neural network to directly generate the output, using the same loss as in [Gatys et al. 2015a]. This setup moves the computational burden to training time, resulting in faster inference time. However, the learned network can only synthesize the texture it was trained on and cannot generalize to new textures.

A more recent line of work [Alanov et al. 2019; Bergmann et al. 2017; Frühstück et al. 2019; Jetchev et al. 2016; Li and Wand 2016; Li et al. 2017b; Shaham et al. 2019; Zhou et al. 2018] has proposed using Generative Adversarial Networks (GANs) for more realistic texture synthesis while still suffering from the inability to generalize to new unseen textures.

Zhou et al. [Zhou et al. 2018] learn a generator network that expands $k \times k$ texture blocks into $2k \times 2k$ output through a combination of adversarial, L_1 , and style (gram matrix) loss. Li et al. and Shaham et al. [Li and Wand 2016; Shaham et al. 2019] use a special discriminator that examines statistics of local patches in feature space. However, even these approaches can only synthesize a single texture which it has been trained on.

Other efforts [Alanov et al. 2019; Bergmann et al. 2017; Frühstück et al. 2019; Jetchev et al. 2016; Li et al. 2017b] try to train on a set of texture images. During test time, the texture being generated is either chosen by the network [Bergmann et al. 2017; Jetchev et al. 2016] or user-controlled [Alanov et al. 2019; Li et al. 2017b]. [Frühstück et al. 2019] propose a non-parametric method to synthesize large-scale, varied outputs by combining intermediate feature maps. However, these approaches limit generation to textures available in the training set, and thus are unable to produce unseen textures out of the training set.

Li et al. [Li et al. 2017a] apply a novel whitening and colorizing transform to an encoder-decoder architecture, allowing them to generalize to unseen textures, but rely on inner SVD decomposition which is slow. Additionally, it can only output texture images with the same size as the input.

Yu et al. [Yu et al. 2019] perform the interpolation between two or more source textures. While forcing two source textures to be identical can convert it to the texture synthesis setting, it will reduce the framework to be more like a conventional CNN. Besides probably suffering from the issues of conventional CNNs, its main operations of per-feature-entry shuffling, retiling and blending would greatly break the regular or large structure patterns in the input.

Other Image-to-Image Tasks. GANs [Goodfellow et al. 2014] have also been used in other image-to-image tasks [Dundar et al. 2020; Isola et al. 2016; Wang et al. 2018]. Ledig et al. [Ledig et al. 2016] used it to tackle the problem of super-resolution, where detail is added to a low-resolution image to produce high-definition output. In contrast to these tasks, the texture synthesis problem is to synthesize new, varied regions similar to the input, and not to provide more details to an existing layout as in [Ledig et al. 2016] or translate the texture to a related domain as in [Isola et al. 2016]. Other recipes like converting texture synthesis to an image inpainting problem usually cannot get satisfying results as they usually cannot handle big holes where we need to do the synthesis.

Similarity Map. Our framework relies on computing the self-similarity map, which is similar in spirit to the deep correlation formulation in [Sendik and Cohen-Or 2017]. The difference is that [Sendik and Cohen-Or 2017] computes a dot product map between the feature map and its spatially shifted version and uses it as a regularizing term in their optimization objective; in contrast, we aggregate all the channels' information to compute a single-channel difference similarity map and use it to model the optimal synthesis scheme in the network with a single pass.

3 OUR APPROACH

Problem Definition: Given an input texture patch, we want to expand the input texture to a larger output whose local pattern resembles the input texture pattern. Our approach to this problem shares some similar spirits with the traditional assembling based methods which try to find the optimal displacements of copies of the input texture, as well as the corresponding assembly scheme to produce a large, realistic texture image output. We will first formulate the texture expansion problem as a weighted linear combination of displaced deep features at various shifting positions, and then discuss how to use the transposed convolution operation to address it.

Deep Feature Expansion: Let $\mathbb{F} \in \mathbb{R}^{C \times H \times W}$ be the deep features of an input texture patch, with C , H and W being the number of channels, the height, and width, respectively. We create a spatially expanded feature map, for instance by a factor of 2, by simply pasting and accumulating \mathbb{F} into a $C \times 2H \times 2W$ space. This is done by shifting \mathbb{F} along the width axis with a progressive step ranging from 0 to W , as well as along the height axis with a step ranging from 0 to H . All the shifted maps are then aggregated together to give us an expanded feature map $\mathbb{G} \in \mathbb{R}^{C \times 2H \times 2W}$.

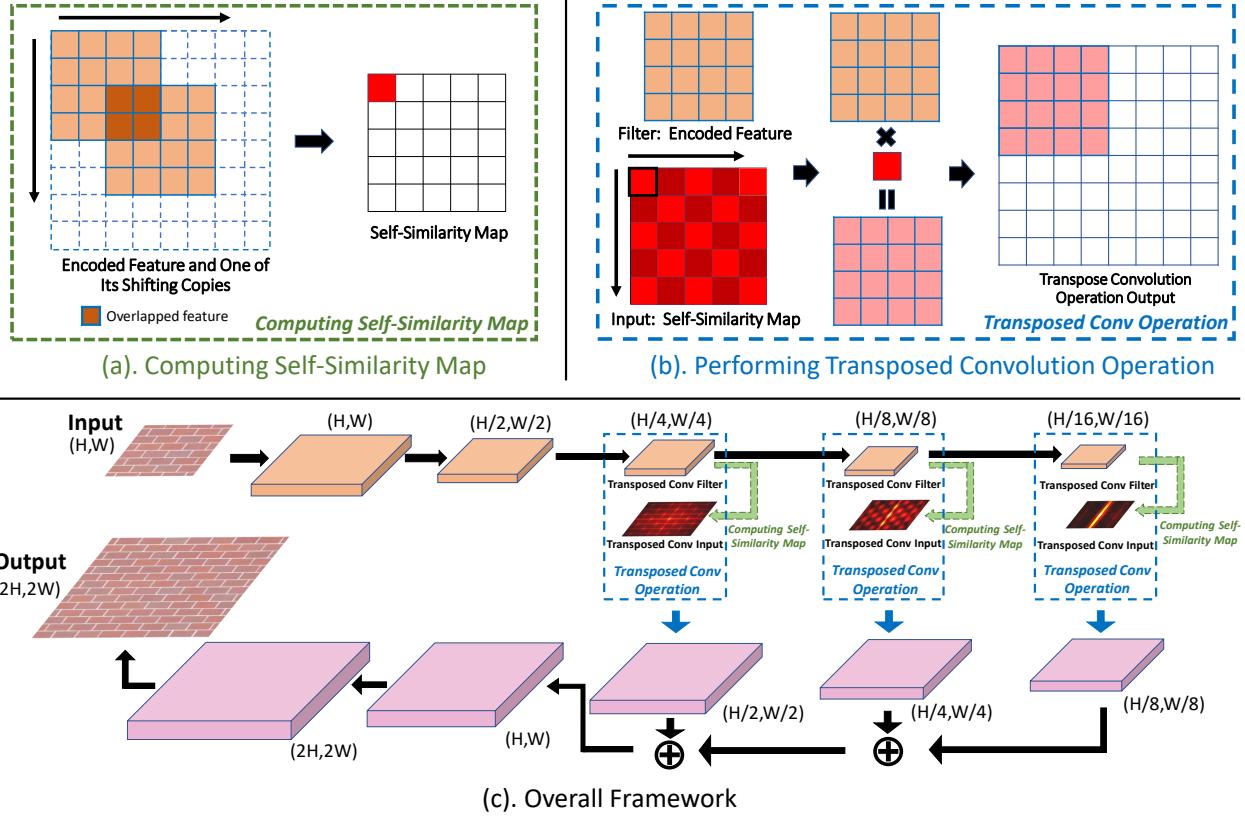


Fig. 2. (a) shows how the self-similarity map is computed. (b) shows how to perform the transposed convolution operation. Both (a) and (b) are the components used in our overall framework (c), shown with green and blue colors, respectively. Full animation of (a) and (b) can be found in the supplementary video. In (c), yellow boxes represent the features in the encoder. The encoded features in the last three scales are first used to compute the self-similarity maps, as shown in (a). We then perform the transposed convolution operation as shown in (b), where encoded features are used as transposed convolution filters to convolve the self-similarity maps. The convolved outputs are then used in the decoder to generate the final image.

To aggregate the shifted copies of \mathbb{F} , we compute a weighted sum of them. For instance, to calculate the feature $G(i, j) \in \mathbb{R}^C$, we aggregate all possible shifted copies of $\mathbb{F}(\cdot, \cdot) \in \mathbb{R}^C$ that fall in the spatial location (i, j) . While previous approaches rely on hand crafted or other heuristics to aggregate the overlapping features, in our approach, we propose to weight each shifted feature map with a similarity score that quantifies the semantic distance between the original \mathbb{F} and its shifted copy. Finally, aggregation is done by simple summation of the weighted features. Mathematically, \mathbb{G} can be given by

$$\mathbb{G}^c = \sum_{p,q} s(p, q) \mathbb{E}_{p,q}^c \quad (1)$$

where $c \in [0, C]$, $i \in [0, 2H]$, $j \in [0, 2W]$, $p \in [-H/2, H/2]$, $q \in [-W/2, W/2]$, $s_{p,q}$ is the similarity score of (p, q) -shifting, and $\mathbb{E}_{p,q}$ is the projection of \mathbb{F} 's (p, q) -shifted copy on the $(2H, 2W)$ grid. Namely, $\mathbb{E}_{p,q}^c(i+p+H/2, j+q+W/2) = \mathbb{F}^c(i, j)$, with $\mathbb{E} \in \mathbb{R}^{C \times 2H \times 2W}$, $\mathbb{F} \in \mathbb{R}^{C \times H \times W}$, $i \in [0, H]$ and $j \in [0, W]$.

We compute the similar score $s(p, q)$ of current (p, q) -shifting using the overlapping region based on the following equation:

$$s(p, q) = -\frac{\sum_{m,n,c} \|\mathbb{F}^c(m, n) - \mathbb{F}^c(m-p, n-q)\|^2}{\sum_{m,n,c} \|\mathbb{F}^c(m, n)\|^2} \quad (2)$$

Here, $m \in [\max(0, p), \min(p+H, H)]$ and $n \in [\max(0, q), \min(q+W, W)]$ indicate the overlapping region between current (p, q) -shifted copy and the original copy. $\|\mathbb{F}\|_2$ is the L2 norm of \mathbb{F} . The dominator $\sum_{m,n,c} \|\mathbb{F}^c(m, n)\|^2$ is used for denormalization such that the scale of $s(p, q)$ is independent of the scale of \mathbb{F} . Figure 2(a) shows how the self-similarity score is computed at shifting $(-H/2, -W/2)$. Note that self-similarity map is not symmetric with respect to its center as the dominator of Equation 2 is not symmetric with respect to the center. Full animation of computing the self-similarity map can be found in the supplementary video. We will apply some simple transformation on s before using it in Equation 1, specifically one convolutional layer and one activation layer in our implementation.

As shown in Equation 2, the similarity score for a shift of (p, q) along the width and height axis, respectively, is calculated as the L2 distance between the *un-shifted* and *shifted* copies of the feature map, normalized by the norm of the un-shifted copy's overlapping region. So, a shift of $(p = 0, q = 0)$ gives the maximum score because there

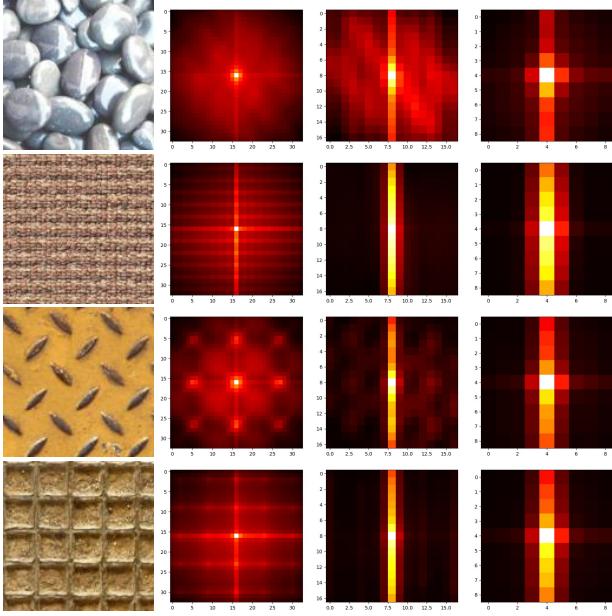


Fig. 3. Self-similarity Maps. We show the input texture images and the visualization of their self-similarity maps at 3 different scales ($1/4 \times 1/4$, $1/8 \times 1/8$, $1/16 \times 1/16$). The first texture image exhibits more obvious self-similarity patterns at the second scale, while other three texture images exhibits more obvious self-similarity patterns at the first scale.

is no shifting and it exactly matches the original copy. Computing self-similarity maps can be efficiently implemented with the help of existing convolution operations. Details are discussed in the supplementary file.

We compute the self-similarity maps at multiple scales. Different texture images may exhibit more obvious self-similarity patterns on a specific scale than other scales, as shown in Figure 3.

Feature (Texture) Expansion via Transposed Convolution Operation: Note that the process of pasting shifted feature maps and aggregating them to create larger feature maps is equivalent to the operation of a standard transposed convolution in deep neural networks. For the given filter and input data, a transposed convolution operation simply copies the filter weighted by the respective center entry’s data value in the input data into a larger feature output grid, and perform a summation. In fact, our proposed Equation 1 is similar with a transposed convolution. Specifically, we apply transposed convolutions with a stride of 1, treating the feature map $\mathbb{F} \in \mathbb{R}^{C \times H \times W}$ as the transposed convolution filter, and the similarity map $\mathbb{S} \in \mathbb{R}^{1 \times (H+1) \times (W+1)}$, given by Equation 2, as the input to the transposed convolution. This results in an output feature map \mathbb{G} of size $C \times 2H \times 2W$. Figure 2(b) shows how the transposed convolution is done using the encoded input texture as filters and the first entry in the self-similarity map as input. Full animation of the transposed convolution operation can be found in the supplementary video.

3.1 Architecture

Figure 2(c) illustrates our overall texture synthesis framework. It relies on a UNet-like architecture. The encoder extracts deep features of the input texture patch at several scales. We then apply our proposed transposed convolution-based feature map expansion technique at each scale. The resulting expanded feature map is then passed onto a standard decoder layer. Our network is fully differentiable, allowing us to train our model with stochastic gradient-based optimizations in an end-to-end manner. The four main components of our framework in Figure 2(c) are:

- (1) **Encoder:** Learns to encode the input texture image into deep features at different scales or levels.
- (2) **Self-Similarity Map Generation:** Constructs guidance maps from the encoded features to weight the shifted feature maps in the shift, paste and aggregate process of feature map expansion (see Equation 2 and Figure 2(a)).
- (3) **Transposed Convolution Operation:** Applies spatially varying transposed convolution operations, treating the encoded feature maps directly as *filters* and the self-similarity maps as *inputs* to produce expanded feature maps, as shown in Figure 2(b). Note that, unlike traditional transposed convolution layers, ours transposed convolution filters are not learning parameters. More details about the difference between our transposed convolution operation and traditional transposed convolution layer can be found in the supplemental file.
- (4) **Decoder:** Given the already expanded features from the transposed convolution operations at different scales, we follow the traditional decoder network design that uses standard convolutional layers followed by bilinear upsampling layers to aggregate features at different scales, and generate the final output texture, as shown in the last row of Figure 2(c).

As described above, our proposed texture expansion technique is performed at multiple feature representation levels, allowing us to capture both diverse features and their optimal aggregation weights. Unlike previous approaches that rely on heuristics or graph-base techniques to identify the optimal overlap of shifted textures, our approach formulates the problem as a direct generation of larger texture images conditioned on optimally assembled deep features at multiple scales. This makes our approach desirable as it is data-driven and generalizable for various textures.

3.2 Loss Functions

During training, given a random image with size $(2H, 2W)$, denoted as I_{target} , its center crop image with size (H, W) will be the input to the network, denoted as I_{input} . We train the network to predict an output image I_{out} with the size $(2H, 2W)$. Both VGG-based perceptual loss, style loss and GAN loss are used to train the network. The perceptual loss and style loss are defined between I_{target} and I_{out} at the full resolution of $(2H, 2W)$; meanwhile, the GAN loss is defined on the random crops at the resolution of (H, W) . Details are discussed below.

VGG-based perceptual loss and style loss. Perceptual loss and style loss are defined following Gatys et al. [Gatys et al. 2015b].

The perceptual loss and style loss are defined as:



Fig. 4. Ablation study on the components of loss functions.

$$\mathcal{L}_{perceptual} = \sum_{p=0}^{P-1} \frac{\|\Psi_p^{\text{I}_{out}} - \Psi_p^{\text{I}_{target}}\|_1}{N_{\Psi_p^{\text{I}_{target}}}} \quad (3)$$

$$\mathcal{L}_{styleout} = \sum_{p=0}^{P-1} \frac{1}{C_p C_p} \left\| K_p ((\Psi_p^{\text{I}_{out}})^\top (\Psi_p^{\text{I}_{out}}) - (\Psi_p^{\text{I}_{target}})^\top (\Psi_p^{\text{I}_{target}})) \right\|_1 \quad (4)$$

Here, $N_{\Psi_p^{\text{I}_{target}}}$ is the number of entries in $\Psi_p^{\text{I}_{target}}$. The perceptual loss computes the L^1 distances between both I_{out} and I_{target} , but after projecting these images into higher level feature spaces using an ImageNet-pretrained VGG-19 [Simonyan and Zisserman 2014]. $\Psi_p^{\text{I}_*}$ is the activation map of the p th selected layer given original input I_* . We use feature from 2-nd, 7-th, 12-th, 21-st and 30-th layers corresponding to the output of the RELU layers at each scale. In Equation (4), the matrix operations assume that the high level features $\Psi(x)_p$ is of shape $(H_p W_p) \times C_p$, resulting in a $C_p \times C_p$ Gram matrix, and K_p is the normalization factor $1/C_p H_p W_p$ for the p th selected layer.

GAN loss. The discriminator takes the concatenation of I_{input} and a random crop of size (H, W) from either I_{out} or I_{target} as input. Denote the random crop from I_{out} as $\text{I}_{randcrop}^{\text{out}}$ and the random crop from I_{target} as $\text{I}_{randcrop}^{\text{target}}$. The intuition of using concatenation is to let the discriminator learn to classify whether I_{input} and $\text{I}_{randcrop}^*$ is a pair of two similar texture patches or not. We randomly crop 10 times for both I_{out} and I_{target} and sum up the losses.

Ablation Study. These 3 losses are summed up with the weights of 0.05, 120 and 0.2 respectively. We find that all of them are useful and necessary. As shown in Figure 4, without perceptual loss, the result just looks like the naive tiling of the inputs; no style loss makes the border region blurry; and no GAN loss leads to obvious discrepancy between the center region and the border region.

4 EXPERIMENTS AND COMPARISONS

4.1 Dataset & Training

To train our network, we collected a large dataset of texture images. We downloaded 55,583 images from 15 different texture image sources [Abdelmounaim and Dong-Chen 2013; Burghouts and Geusebroek 2009; Center for Machine Vision Research [n.d.]; Cimpoi et al. 2014; Dai et al. 2014; Fritz et al. 2004; Mallikarjuna et al. 2006; Picard et al. 2010; ?]. The total dataset consists of texture images with a wide variety of patterns, scales, and resolutions. We

randomly split the dataset to create a training set of 49,583 images, a validation set of 1,000 images, and a test set of 5,000 images. All generation results and evaluation results in the paper are from the test set. When using these images, we resize them to the target output size as the ground truth and the center cropping of it as input.

Our network utilizing the transposed convolution operation is implemented using the existing PyTorch interface without custom CUDA kernels. We trained our model on 4 DGX-1 stations with 32 total NVIDIA Tesla V100 GPUs using synchronized batch normalization layers [Ioffe and Szegedy 2015]. For 128×128 to 256×256 synthesis, we use batch size 8 and trained for 600 epochs. The learning rate is set to be 0.0032 at the beginning and decreased by 1/10 every 150 epochs. For 256×256 to 512×512 synthesis, we fine-tuned the model based on the pre-trained one for 128×128 to 256×256 synthesis for 200 epochs. While directly using 128 to 256 synthesis pre-trained model generates reasonable results, fine-tuning leads to better quality.

4.2 Baseline & Evaluation Metrics

Baselines. We compare against several baselines: 1) **Naive tiling** which simply tiles the input four times; 2) **Self-tuning** [Kaspar et al. 2015], the state-of-the-art optimization-based method; 3) **pix2pixHD** [Wang et al. 2018], the state-of-the-art image-to-image translation network where we add one more upsampling layer to generate an output 2x2 larger than the input; 4) **WCT** [Li et al. 2017a] is the style transfer method; 5) **DeepTexture** [Gatys et al. 2015a], an optimization based using network features, for which we directly feed the ground truth as input; 6) **Texture Mixer** [Yu et al. 2019], a texture interpolation method where we set the interpolation source patches to be all from the input texture; 7) **Non-stationary (Non-stat.)** [Zhou et al. 2018] and **SinGAN** [Shaham et al. 2019], both of which overfit one model per texture. We train **Non-stat.** and **SinGAN** for two versions respectively; one version with direct access to the exact ground truth at the exact target size, and one version without access to target-size ground truth but only the input. In the paper, * will correspond to methods that either directly take ground truth images for processing or are overfitting the model to ground truth.

Table 1 shows the runtime and corresponding properties for all the methods. Compared with **Self-tuning**, our method is much faster. In contrast to **Non-stat.** and **SinGAN**, transposer (ours) generalizes better and hence does not require per image training. Comparing with **DeepTexture** and the style transfer method **WCT**, our method

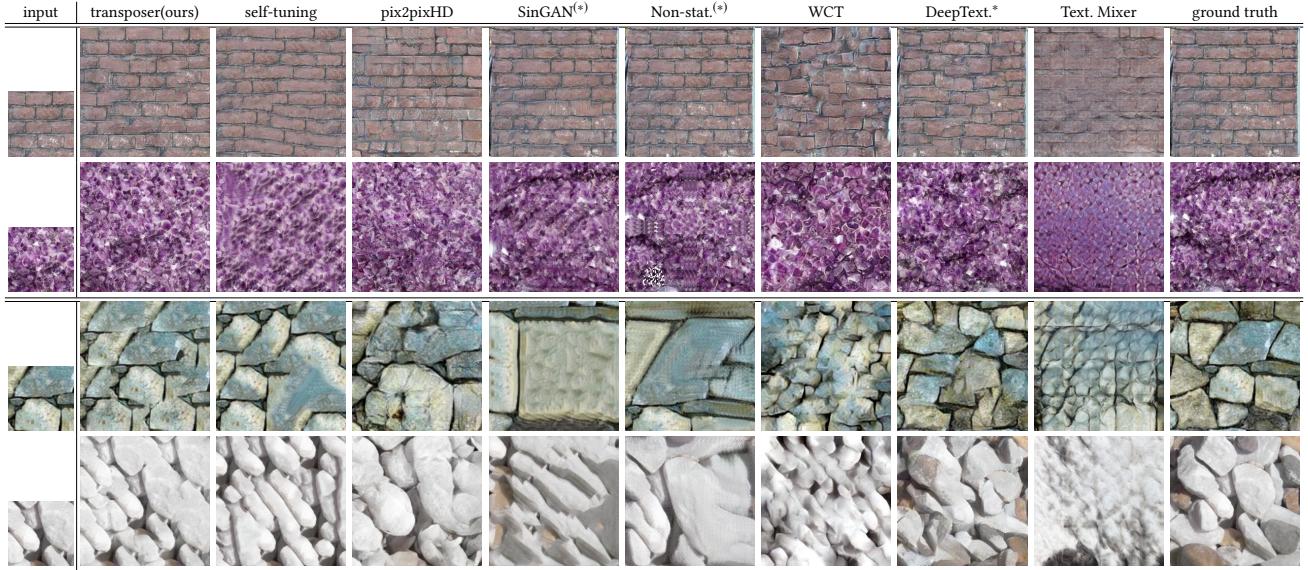


Fig. 5. Results of different approaches on 128×128 to 256×256 texture synthesis. For SinGAN(*) and Non-stat.(*) results, the first two rows show the results when training with direct access to exact-size ground truth; the remaining 2 rows show the results without them accessing the ground truth. In this paper, unless specified, our results (transposer) uses self-similarity map as transposed convolution inputs by default.

Method	Time		Properties	
	256x256	512x512	Generalizability	Size-increasing
Self-tuning[Kaspar et al. 2015]	140 s	195 s	Good	Yes
Non-stationary[Zhou et al. 2018]	362 mins	380 mins	No	Yes
SinGAN[Shaham et al. 2019]	45 mins	100 mins	No	Yes
DeepTexture[Gatys et al. 2015a]	13 mins	54 mins	No	No
WCT[Li et al. 2017a]	7 s	14 s	Medium	Yes
pix2pixHD [Wang et al. 2018]	11 ms	22 ms	Medium	Yes
Texture Mixer [Yu et al. 2019]	-	799 ms	Medium	Yes
transposer(ours)	43 ms	260 ms	Good	Yes

Table 1. Time required for synthesis at different spatial resolutions for various approaches and their corresponding properties. For Non-stationary and SinGAN, the reported time includes training time. All methods are run on one NVIDIA Tesla V100, except for Self-tuning which runs the default 8 threads in parallel on an Intel Core i7-6800K CPU @ 3.40GHz.

is still much faster without the need of iterative optimization or SVD decomposition. Even though **pix2pixHD** is faster than our method, it cannot perform proper texture synthesis as shown in Figures 5 and 6, same as Texture Mixer [Yu et al. 2019].

Evaluation Metrics. To the best of our knowledge, there is no standard metric to quantitatively evaluate texture synthesis results. We use 3 groups of metrics (6 in total):

- (1) **Existing metrics** include SSIM [Wang et al. 2004], Learning Perceptual Image Patch Similarity (LPIPS) [Zhang et al. 2018] and Fréchet Inception Distance (FID) [Heusel et al. 2017]. SSIM and LPIPS are evaluated using image pairs. FID measures the distribution distance between the generated image set and the ground truth image set in feature space.
- (2) **Crop-based metrics designed for texture synthesis evaluation** include crop-based LPIPS (c-LPIPS) and crop-based FID (c-FID). While the original LPIPS and FID are computed

on full-size images, c-LPIPS and c-FID operate on crops of images. For c-FID, we crop a set of images from the output image and crop the other set from the ground truth image, and then compute the FID between these two sets (we use a dimension of 64 for c-FID instead of the default 2048 due to a much smaller image set). For c-LPIPS, we compute the LPIPS between the input image and one of the 8 random crops from the output image, and average the scores among the 8 crops.

- (3) **User Study.** Another way to measure the performances of different methods is by performing user study. We thus use Amazon Mechanical Turk (AMT) to evaluate the quality of synthesized textures. We perform AB tests where we provide the user the input texture image and two synthesized images from different methods. We then ask users to choose the one with better quality. Each image is viewed by 20 workers, and the orders are randomized. The obtained preference scores

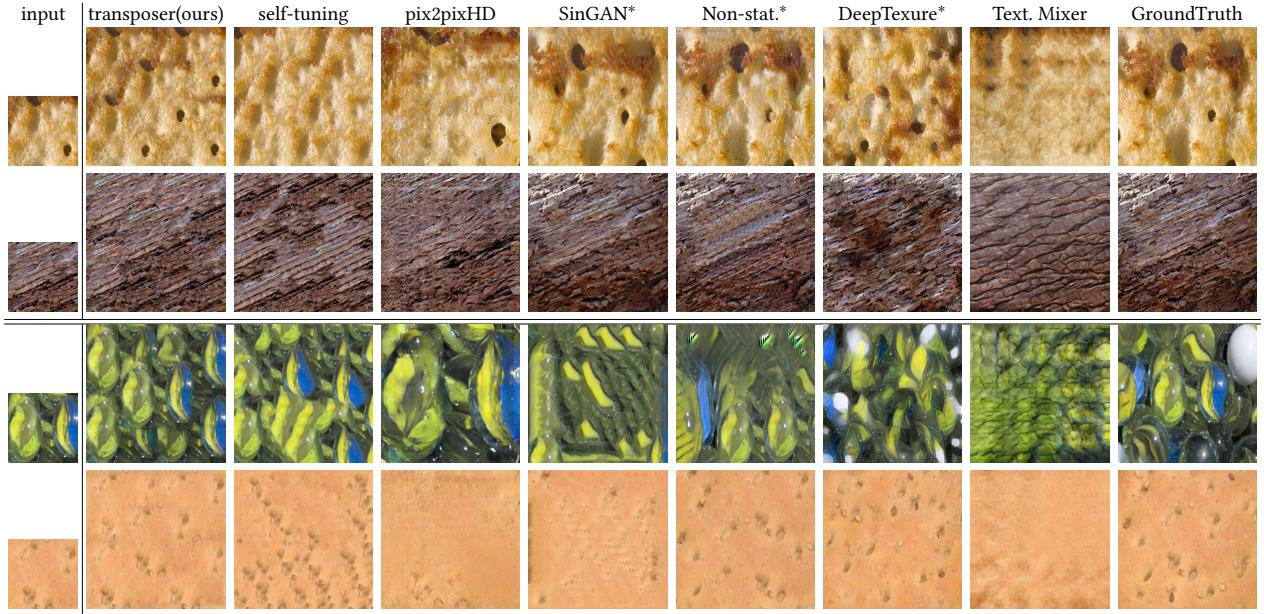


Fig. 6. Results of different approaches on 256×256 to 512×512 texture synthesis. For SinGAN(*) and Non-stat.(*) results, the first two rows show the results when training with direct access to exact-size ground truth; the remaining 2 rows show the results without them accessing the ground truth.

	SSIM	FID	c-FID	LPIPS	c-LPIPS
Naive tiling	0.311	23.868	0.5959	0.3470	0.2841
Self-tuning	0.3075	33.151	0.5118	0.3641	0.2970
pix2pixHD	0.318	26.800	0.5687	0.3425	0.2833
WCT	0.280	57.630	0.4347	0.3775	0.3226
transposer (ours)	0.386	21.615	0.4763	0.2709	0.2653
Ground Truth*	1	0	0.1132	0	0.2670

Table 2. Synthesis scores for different approaches averaged over 5,000 images.

(Pref.) are shown in Table 3, which indicate the portion of workers that prefer our result over the other method.

4.3 Comparison Results

4.3.1 Evaluating synthesis of size 128 to 256. We compare with Self-tuning, pix2pixHD and WCT on the whole test set of 5,000 images and show the quantitative comparisons in Table 2. It is noticeable that our method outperforms Self-tuning and pix2pixHD for all the metrics.

Due to the fact that Non-stat., SinGAN and DeepTexture are too slow to evaluate on all 5,000 test images, we randomly sampled 200 from the 5,000 test images to evaluate them. The visual comparison is shown in Figure 5. The numerical evaluation results are summarized in Table 3. As shown in 2nd-8th rows of Table 3, our method significantly out-performs all the methods which do not directly take ground truth as input. When compared with Self-tuning, we achieve better LPIPS score (0.273 vs. 0.358), and 63% of people prefer the results generated by our method over the ones generated by Self-tuning. The remaining rows of Table 3 also show that our method

	SSIM	FID	c-FID	LPIPS	c-LPIPS	Pref.
Naive tiling	0.289	77.54	0.552	0.349	0.287	-
Self-tuning	0.296	101.75	0.464	0.358	0.292	0.63
Non-stat.	0.321	143.31	2.728	0.3983	0.3436	0.92
SinGAN	0.337	212.30	1.375	0.3924	0.3245	0.81
pix2pixHD	0.299	93.70	0.456	0.354	0.292	0.66
WCT	0.280	126.10	0.401	0.375	0.300	0.67
Texture Mixer	0.311	211.78	1.997	0.399	0.334	0.89
transposer(ours)	0.437	74.35	0.366	0.273	0.272	
Ground Truth*	1	0	0.112	0	0.270	0.51
Non-stat.*	0.767	73.72	2.149	0.1695	0.3276	-
SinGAN*	0.492	88.14	1.137	0.2467	0.2939	-
DeepTexture*	0.289	67.89	0.289	0.336	0.298	0.46

Table 3. Synthesis scores for different approaches averaged over 200 images.

performs better than other size-increasing baselines (Non-stat.* and SinGAN*) and performs better or similar to DeepTexture*, which all take ground truth as input. For instance, 51% people prefer our results over the ground truth and 46% of people prefer our results over DeepTexture, which directly takes ground truth for its optimization.

4.3.2 Evaluating synthesis of size 256 to 512. We also evaluate on 256×256 image to 512×512 image synthesis using the same metrics. We show the quantitative results in the supplementary file. Visual comparisons can be found in Figure 6. It confirms that our approach produces superior results. For example, Self-tuning almost completely misses the holes in the 1st input texture image, and pix2pixHD simply enlarges the local contents instead of performing synthesis. In Figure 7, we show the 4×4 times larger texture synthesis results using our framework. This is done by running the



Fig. 7. More of our results on 4×4 times larger synthesis. Zoom in for more details.

	SSIM	FID	c-FID	LPIPS	c-LPIPS
Self-sim. Map (default)	0.437	74.35	0.366	00.273	0.272
Learnable TransConv	0.3087	88.05	0.387	0.331	0.2797
Fixed Map	0.2966	97.79	0.383	0.3554	0.2848
Random Map	0.2959	76.51	0.387	0.336	0.2645

Table 4. Ablation study for transposed convolution operation and self-similarity map. For SSIM, the higher the better; for other metrics, the lower the better. The first row represents the transposer framework taking self-similarity map as inputs, the default setting in this paper.

transformer network twice with each performing 2×2 times larger synthesis.

4.4 Ablation Study and Random Noise as Input

4.4.1 Ablation study. To understand the role of self-similarity map, we conduct three additional ablation study experiments: 1). **Learnable TransConv**: using the traditional transposed convolution layer with learnable parameters instead of directly using encoded feature as filters and its self-similarity map as input, while keeping other network parts and training strategies unchanged; 2). **Fixed Map**: using fixed sampled maps instead of self-similarity maps;

3). **Random Map**: using randomly sampled maps instead of self-similarity maps. As shown in Figure 2, we have 3 different scales' features, for running **Fixed map** and **Random map**, we sample the map for the smallest scale's feature and then bilinear upsampling it for the other two scales. Table 4 and Figure 8 show the quantitative and qualitative results, respectively. These 3 settings are compared with the default transformer setting, using self-similarity map as transposed convolution input. It can be seen that **Learnable TransConv** with the traditional learnable transposed convolution layer will simply enlarge the input rather than perform reasonable synthesis, similar to pix2pixHD. This confirms our hypothesis that conventional CNN designs with traditional (de)convolution layers and up/down-sampling layers cannot capture the long-term structural dependency required by texture synthesis. **Fixed map** can't produce faithful results. On the other hand, using random noise map as transposed convolution input has both advantages and disadvantages, as discussed below.

4.4.2 Trade-off between self-similarity map and random noise map. In the last column of Figure 8, the 1st row shows that sampling a random noise map at test time can successfully generate diverse

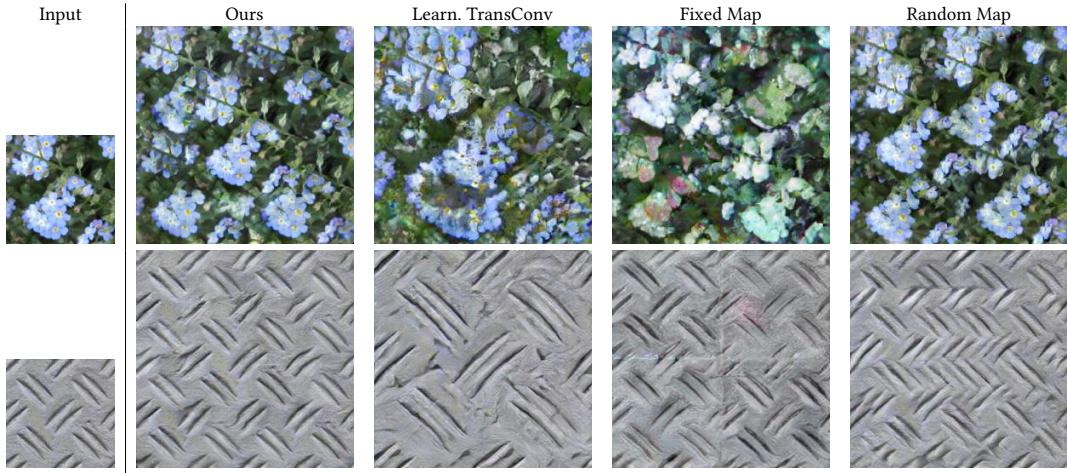


Fig. 8. Ablation study for using transposed convolution operations and self-similarity maps. It can be seen that without using them, the results become much worse.



Fig. 9. Direct 2048×2048 texture generation from 128×128 input by sampling random noise maps. Zoom in for more details. Left small image is the input; right large image is the output.

results. However, note that the self-similarity map is critical in identifying the structural patterns and preserving them in the output. In the 2nd row of Figure 8, the result of using self-similarity maps successfully preserved the regular structures, while using random noise maps failed. We believe that in practice, there is a trade-off between preserving the structure and generating variety. For input texture images with regular structural patterns, self-similarity map provides better guidance for the transposed convolution operation to preserve these structural patterns. On the other hand, using random noise map as inputs can generate diverse outputs by sampling

different noise maps, shown in Figure 10 and it is also possible to directly generate arbitrary large texture outputs by sampling larger noise map, shown in Figure 9 while using self-similarity map can only do smaller than 3×3 times larger synthesis, limited by the size of self-similarity map.

5 CONCLUSION & DISCUSSION

In this paper, we present a new deep learning based texture synthesis framework built based on transposed convolution operations. In our framework, the transposed convolution filter is the encoded

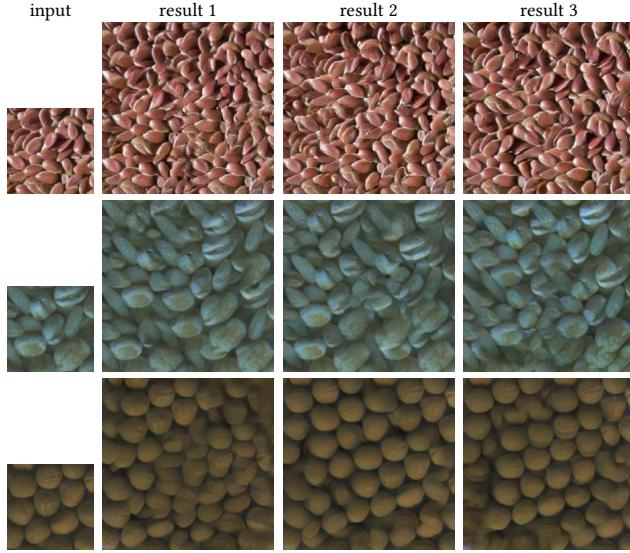


Fig. 10. Diverse outputs given different random noises as inputs.

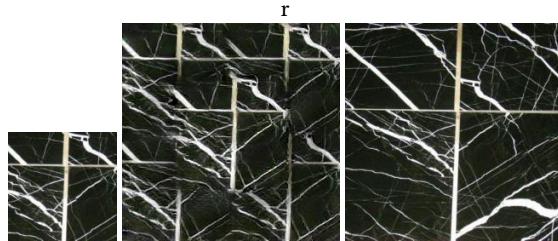


Fig. 11. Failure case for our method. From left to right: input, our synthesis result and the ground truth.

features of the input texture image, and the input to the transposed convolution is the self-similarity map computed on the corresponding encoded features. Quantitative comparisons based on existing metrics, our specifically designed metrics for texture synthesis, and user study results all show that our method significantly outperforms existing methods, while our method also being much faster. Self-similarity map helps preserve the structure better while random noise map allows to generate diverse results. Some further research could also be providing more controlable flexibility by combining both self-similarity map and random noise map as inputs. One limitation of our method is that it fails to handle sparse thin structures like shown in Figure 11 and highly non-stationary inputs [Zhou et al. 2018]. As some highly non-stationary textures mainly emphasize the effect on some specific direction, one possible solution to deal with them may be emphasizing the similarity score on specific directions while suppressing it on other directions to capture directional effects, and/or using cropped, resized or rotated feature maps as transposed convolution filters to capture the effects of textons repeating with various forms. We leave these as future research exploration. While existing deep learning-based image synthesis methods mostly focus on taking the inputs from other modalities like semantic maps or

edge maps, we believe our method will also stimulate more deep learning researches for exemplar-based synthesis.

ACKNOWLEDGMENTS

We would like to thank Brandon Rowllet, Sifei Liu, Aysegul Dundar, Kevin Shih, Rafael Valle and Robert Pottoroff for valuable discussions and proof-reading.

REFERENCES

- Safia Abdelmounaima and He Dong-Chen. 2013. New Brodatz-Based Image Databases for Grayscale Color and Multiband Texture Analysis. *ISRN Machine Vision*.
- Aibek Alanov, Max Kochurov, Denis Volkhonskiy, Daniil Yashkov, Evgeny Burnaev, and Dmitry Vetrov. 2019. User-Controllable Multi-Texture Synthesis with Generative Adversarial Networks. *arXiv preprint arXiv:1904.04751* (2019).
- Urs Bergmann, Nikolay Jetchev, and Roland Vollgraf. 2017. Learning texture manifolds with the periodic spatial GAN. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR.org, 469–477.
- Gertjan J. Burghouts and Jan-Mark Geusebroek. 2009. Material-specific adaptation of color invariant features. *Pattern Recognition Letters* 30, 3 (2009), 306 – 313. <https://doi.org/10.1016/j.patrec.2008.10.005>
- University of Oulu Finland Center for Machine Vision Research. [n.d.]. Outex Texture Database. ([n. d.]). <http://www.outex.oulu.fi/>
- M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. 2014. Describing Textures in the Wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- D. Dai, H. Riemenschneider, and L. Van Gool. 2014. The Synthesizability of Texture Examples. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan Goldman, and Pradeep Sen. 2012. Image melding: combining inconsistent images using patch-based synthesis. *ACM Transactions on Graphics (ToG)* 31, 4 (2012), 1–10.
- Vincent Dumoulin and Francesco Visin. 2016. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285* (2016).
- Aysegul Dundar, Karan Sapra, Guilin Liu, Andrew Tao, and Bryan Catanzaro. 2020. Panoptic-based Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8070–8079.
- Alexei A Efros and William T Freeman. 2001. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 341–346.
- Alexei A Efros and Thomas K Leung. 1999. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, Vol. 2. IEEE, 1033–1038.
- Mario Fritz, Eric Hayman, Barbara Caputo, and Jan-Olof Eklundh. 2004. THE KTH-TIPS database.
- Anna Fröhstück, Ibraheem Alhashim, and Peter Wonka. 2019. TileGAN: Synthesis of Large-Scale Non-Homogeneous Textures. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 38, 4 (2019), 58:1–58:11.
- Leon Gatys, Alexander S Ecker, and Matthias Bethge. 2015a. Texture synthesis using convolutional neural networks. In *Advances in neural information processing systems*. 262–270.
- Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015b. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.
- Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. 2001. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 327–340.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium.
- Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).
- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2016. Image-to-Image Translation with Conditional Adversarial Networks. *CoRR abs/1611.07004* (2016). [arXiv:1611.07004](https://arxiv.org/abs/1611.07004) <http://arxiv.org/abs/1611.07004>
- Nikolay Jetchev, Urs Bergmann, and Roland Vollgraf. 2016. Texture Synthesis with Spatial Generative Adversarial Networks. *CoRR abs/1611.08207* (2016). [arXiv:1611.08207](https://arxiv.org/abs/1611.08207) <http://arxiv.org/abs/1611.08207>
- Justin Johnson, Alexandre Alahi, and Fei-Fei Li. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. *CoRR abs/1603.08155* (2016). [arXiv:1603.08155](https://arxiv.org/abs/1603.08155) <http://arxiv.org/abs/1603.08155>

- Alexandre Kaspar, Boris Neubert, Dani Lischinski, Mark Pauly, and Johannes Kopf. 2015. Self tuning texture optimization. In *Computer Graphics Forum*, Vol. 34. Wiley Online Library, 349–359.
- Vivek Kwatra, Irfan Essa, Aaron Bobick, and Nipun Kwatra. 2005. Texture Optimization for Example-based Synthesis. In *ACM SIGGRAPH 2005 Papers* (Los Angeles, California) (*SIGGRAPH '05*). ACM, New York, NY, USA, 795–802. <https://doi.org/10.1145/1186822.1073263>
- Vivek Kwatra, Arno Schödl, Irfan Essa, Greg Turk, and Aaron Bobick. 2003. Graphcut textures: image and video synthesis using graph cuts. In *ACM Transactions on Graphics (ToG)*, Vol. 22. ACM, 277–286.
- Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2016. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *CoRR* abs/1609.04802 (2016). arXiv:1609.04802 <http://arxiv.org/abs/1609.04802>
- Sylvain Lefebvre and Hugues Hoppe. 2006. Appearance-space texture synthesis. *ACM Transactions on Graphics (TOG)* 25, 3 (2006), 541–548.
- Chuan Li and Michael Wand. 2016. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. *CoRR* abs/1604.04382 (2016). arXiv:1604.04382 <http://arxiv.org/abs/1604.04382>
- Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017a. Universal Style Transfer via Feature Transforms. *CoRR* abs/1705.08086 (2017). arXiv:1705.08086 <http://arxiv.org/abs/1705.08086>
- Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017b. Diversified texture synthesis with feed-forward networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3920–3928.
- Lin Liang, Ce Liu, Ying-Qing Xu, Baining Guo, and Heung-Yeung Shum. 2001. Real-time texture synthesis by patch-based sampling. *ACM Transactions on Graphics (ToG)* 20, 3 (2001), 127–150.
- Gang Liu, Yann Gousseau, and Gui-Song Xia. 2016. Texture Synthesis Through Convolutional Neural Networks and Spectrum Constraints. *CoRR* abs/1605.01141 (2016). arXiv:1605.01141 <http://arxiv.org/abs/1605.01141>
- Guilin Liu, Kevin J Shih, Ting-Chun Wang, Fitzsum A Reda, Karan Sapra, Zhiding Yu, Andrew Tao, and Bryan Catanzaro. 2018. Partial Convolution based Padding. *arXiv preprint arXiv:1811.11718* (2018).
- P Mallikarjuna, Alireza Targhi, Mario Fritz, Eric Hayman, Barbara Caputo, and J.-O Eklundh. 2006. THE KTH-TIPS2 database. (07 2006).
- Rosalind Picard, C. Graczyk, Steve Mann, J. Wachman, L. Picard, and L. Campbell. 2010. VisTex vision texture database. (01 2010).
- Javier Portilla and Eero P. Simoncelli. 2000. A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision* 40, 1 (01 Oct 2000), 49–70. <https://doi.org/10.1023/A:1026553619983>
- Yael Pritch, Eitan Kav-Venaki, and Shmuel Peleg. [n.d.] Shift-map image editing. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 151–158.
- Amir Rosenberger, Daniel Cohen-Or, and Dani Lischinski. 2009. Layered Shape Synthesis: Automatic Generation of Control Maps for Non-stationary Textures. In *ACM SIGGRAPH Asia 2009 Papers* (Yokohama, Japan) (*SIGGRAPH Asia '09*). ACM, New York, NY, USA, Article 107, 9 pages. <https://doi.org/10.1145/1661412.1618453>
- Omry Sendik and Daniel Cohen-Or. 2017. Deep correlations for texture synthesis. *ACM Transactions on Graphics (TOG)* 36, 5 (2017), 161.
- Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. 2019. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE International Conference on Computer Vision*, 4570–4580.
- Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- Dmitry Ulyanov, Vadim Lebedev, Victor Lempitsky, et al. 2016. Texture Networks: Feed-forward Synthesis of Textures and Stylized Images. In *International Conference on Machine Learning*, 1349–1357.
- Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.
- Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
- Li-Yi Wei, Sylvain Lefebvre, Vivek Kwatra, and Greg Turk. 2009. State of the Art in Example-based Texture Synthesis. In *Eurographics '09 State of the Art Reports (STARs)*. Eurographics Association. <http://www-sop.inria.fr/reves/Basilic/2009/WLKT09>
- Li-Yi Wei and Marc Levoy. 2000. Fast texture synthesis using tree-structured vector quantization. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 479–488.
- Qing Wu and Yizhou Yu. 2004. Feature matching and deformation for texture synthesis. *ACM Transactions on Graphics (TOG)* 23, 3 (2004), 364–367.
- Ruobing Wu, Wenping Wang, and Yizhou Yu. 2013. Optimized synthesis of art patterns and layered textures. *IEEE transactions on visualization and computer graphics* 20, 3 (2013), 436–446.
- Ning Yu, Connelly Barnes, Eli Shechtman, Sohrab Amirghodsi, and Michal Lukac. 2019. Texture Mixer: A Network for Controllable Synthesis and Interpolation of Texture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12164–12173.
- Jingdan Zhang, Kun Zhou, Luiz Velho, Baining Guo, and Heung-Yeung Shum. 2003. Synthesis of progressively-variant textures on arbitrary surfaces. *ACM Transactions on Graphics (TOG)* 22, 3 (2003), 295–302.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.
- Yang Zhou, Zhen Zhu, Xiang Bai, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2018. Non-stationary Texture Synthesis by Adversarial Expansion. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 37, 4 (2018), 49:1–49:13.

A FRAMEWORK DETAIL

A.1 Self-Similarity Computing & Transposed Convolution

The reviewers are welcome to check the attached animation video showing how a self-similarity map is computed and how the transposed convolution operation is performed.

A.2 Implementation Details for Computing Self-similarity Map

Computing self-similarity map can be efficiently implemented with the help of standard convolution operations. The formula for computing self-similarity map can be relaxed as the following:

$$s(p, q) = -\frac{\sum_{m,n,c} \|\mathbb{F}^c(m, n) - \mathbb{F}^c(m - p, n - q)\|^2}{\sum_{m,n,c} \|\mathbb{F}^c(m, n)\|^2} = -\frac{\sum_{m,n,c} \|\mathbb{F}^c(m, n)\|^2 - 2 * \sum_{m,n,c} \mathbb{F}^c(m, n) * \mathbb{F}^c(m - p, n - q) + \sum_{m,n,c} \|\mathbb{F}^c(m - p, n - q)\|^2}{\sum_{m,n,c} \|\mathbb{F}^c(m, n)\|^2} \quad (5)$$

Here, $m \in [\max(0, p), \min(p + H, H)]$ and $n \in [\max(0, q), \min(q + W, W)]$ indicate the overlapping region between current (p, q) -shifted copy and the original copy. $\|\mathbb{F}\|_2$ is the L2 norm of \mathbb{F} . The dominator $\sum_{m,n,c} \|\mathbb{F}^c(m, n)\|^2$ is used for denormalization such that the scale of $s(p, q)$ is independent of the scale of \mathbb{F} .

Implementation Details. $\sum_{m,n,c} \|\mathbb{F}^c(m, n)\|^2$ can be computed by using \mathbb{F}^2 as convolution input and a convolution filter with weights being all 1s and biases being all 0s. $\sum_{m,n,c} \mathbb{F}^c(m, n) * \mathbb{F}^c(m - p, n - q)$ can be computed by using the zero-padded \mathbb{F} , with $H/2$ zero padding on top and bottom sides and $W/2$ zero padding on left and right sides as convolution input and \mathbb{F} as convolution filter. Similarly, $\sum_{m,n,c} \|\mathbb{F}^c(m - p, n - q)\|^2$ can be computed by using a $(2H, 2W)$ map, with the center region $[H/2 : H/2 + H, W/2 : W/2 + W]$ being 1 and other region being 0, as convolution input and \mathbb{F} as convolution filter.

A.3 Transposed Convolution Block

Table 5 lists the main differences between typical transposed convolution operation and our transposed convolution operation. Fig. 12 shows the details for transposed convolution block in our framework.

typical transposed conv operation		our transposed conv operation
input	output from previous layer	self-similarity map from encoded features
filter	learnable parameters	feature maps from encoder
bias term	learnable parameters	avg-pooling of encoded features with linear transform
filter size	small (e.g. 4x4, 3x3)	large (e.g. 8x8, 16x16, 32x32, 64x64)
stride	2(for upsampling purpose)	1

Table 5. Main differences between typical transposed convolution and our transposed convolution operation

A.4 Network Details

Table 7 shows the details of generator. The discriminator network is the same with pix2pixHD [Wang et al. 2018]. We use partial convolution based padding [Liu et al. 2018] instead of zero padding for all the convolution layers.

	SSIM	FID	c-FID	LPIPS	c-LPIPS
Self-tuning	0.3157	95.829	0.4393	0.4078	0.3653
Non-station.	0.3349	120.245	1.6888	0.4226	0.3911
sinGAN	0.3270	147.9333	1.3806	0.4230	0.3829
pix2pixHD	0.3253	131.655	0.5472	0.4193	0.3780
Ours	0.4533	78.4808	0.3973	0.3246	0.3563
Non-station.*	0.4915	211.0645	1.4274	0.3411	0.3893
sinGAN*	0.2913	154.651	1.6909	0.4787	0.4364
DeepTexture	0.3011	82.053	0.5649	0.4175	0.3830
WCT	0.3124	144.208	0.4125	0.4427	0.4068

Table 6. 256 to 512 synthesis scores for different approaches averaged over 200 images. Non-station.* , sinGAN*, DeepTexture and WCT directly take the ground truth images as inputs.

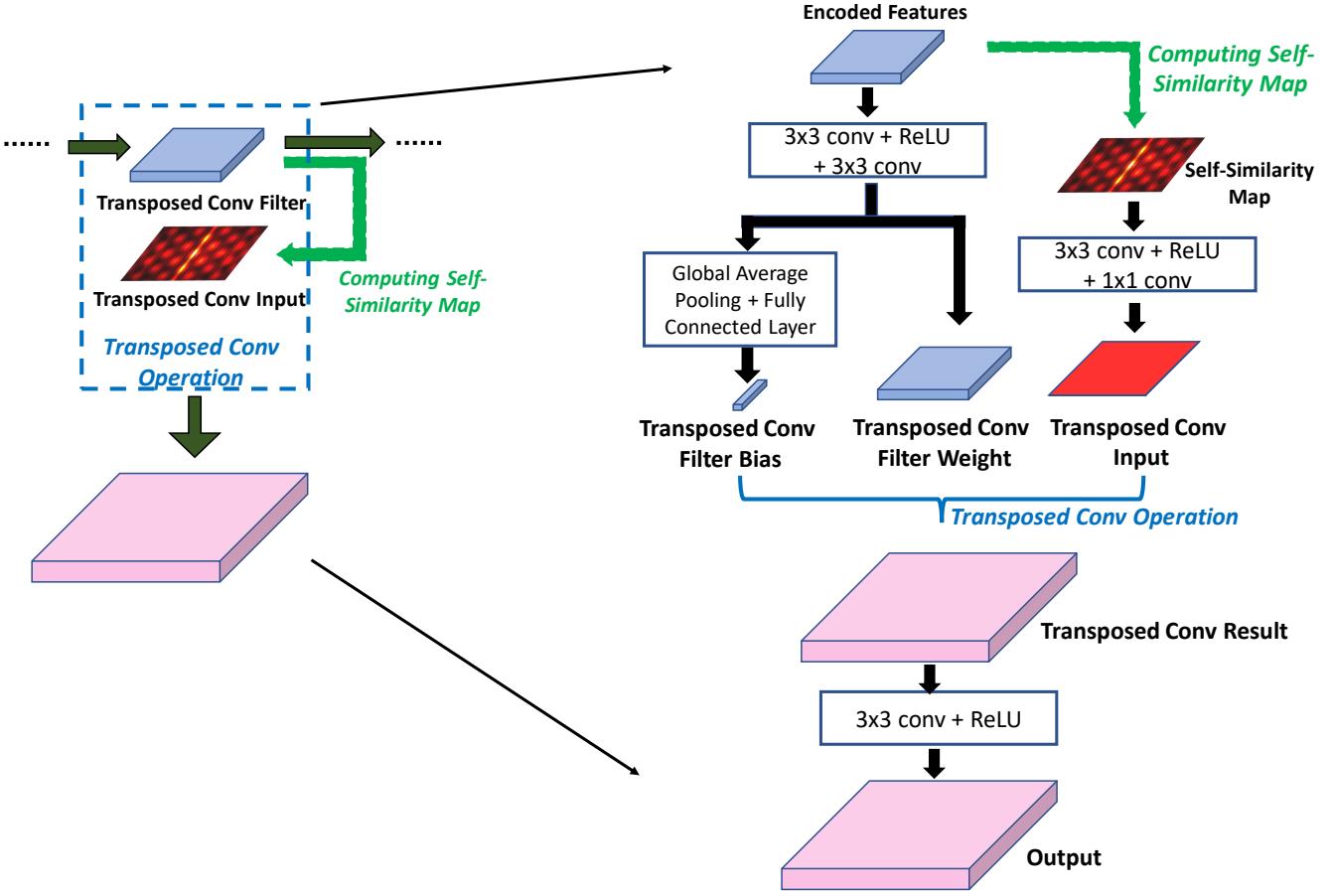


Fig. 12. The details of a transposed convolution block. The left part shows the corresponding preview which is used in the Figure 3 in the main paper; the right part shows the details of this transposed convolution block.

B ADDITIONAL COMPARISON

B.1 256 to 512 Synthesis

In Table 6, we provide the quantitative comparisons for the synthesis results of 256 to 512.

B.2 128 to 256 Synthesis

Non-stat. and Non-stat.* baselines: we take the original code from the author's github repository. The original training strategy for each training iteration is: 1). randomly crop a $2H \times 2W$ from the original big image($> 2H \times 2W$) as the target image; 2). from the target image, randomly crop a $H \times W$ image as the input image. Thus, for 128 to 256 synthesis, to train Non-stat. (without seeing ground truth 256×256 image), for each training iteration, we randomly crop a 96×96 image from the input 128×128 image as target image then from the target image, we randomly crop a 48×48 image as input. To train Non-stat.* (with directly seeing ground truth 256×256 image), for each training iteration, we randomly crop a 128×128 image from the ground truth 256×256 image as the target image and then from the target image, we randomly crop a 64×64 image as input image. For both Non-stat. and Non-stat.* , the inference stage will take 128×128 image as input.

sinGAN and sinGAN* baselines: for training with sinGAN, we used the original author's implementation available on github. And we used the default settings the author provided in their source code. sinGAN code can synthesize textures in two different modes, one that generates a random variation which is of the same size as input texture (we directly using ground truth 256×256 for training, denoted as sinGAN*), and another that generates a texture of larger size (only using 128×128 image, denoted as sinGAN).

	Block	Filter Size	# Filters	Stride/Up Factor	Sync BN	ReLU
Encoder	Conv1	3x3	3 → 64	1	Y	Y
	Conv2_1	3x3	64 → 128	2	Y	Y
	Conv2_2	3x3	128 → 128	1	Y	Y
	Conv3_1	3x3	128 → 256	2	Y	Y
	Conv3_2	3x3	256 → 256	1	Y	Y
	Conv4_1	3x3	256 → 512	2	Y	Y
	Conv4_2	3x3	512 → 512	1	Y	Y
	Conv5_1	3x3	512 → 1024	2	Y	Y
	Conv5_2	3x3	1024 → 1024	1	Y	Y
TransConv_Block3 (w/ Conv3_2)	FilterBranch_Conv1	3x3	256 → 256	1	-	Y
	FilterBranch_Conv2	3x3	256 → 256	1	-	-
	FilterBranch_FC1	-	256 → 256	1	-	-
	SelfSimilarityMapBranch_Conv1	3x3	1 → 8	1	-	Y
	SelfSimilarityMapBranch_Conv2	3x3	8 → 1	1	-	-
TransConv_Block4 (w/ Conv4_2)	transposed Convolution Operation	$\frac{\text{orig_H}}{4} \times \frac{\text{orig_W}}{4}$	filter: 256, input: 1 → 256	-	-	-
	OutputBranch_Conv	3x3	256 → 256	1	-	Y
	FilterBranch_Conv1	3x3	512 → 512	1	-	Y
	FilterBranch_Conv2	3x3	512 → 512	1	-	-
	FilterBranch_FC1	-	512 → 512	1	-	-
TransConv_Block5 (w/ Conv5_2)	SelfSimilarityMapBranch_Conv1	3x3	1 → 8	1	-	Y
	SelfSimilarityMapBranch_Conv2	3x3	8 → 1	1	-	-
	transposed Convolution Operation	$\frac{\text{orig_H}}{8} \times \frac{\text{orig_W}}{8}$	filter: 512, input: 1 → 512	-	-	-
	OutputBranch_Conv	3x3	512 → 512	1	-	Y
Decoder	BilinearUpSample1(w/ TransConv_Block5 output)	-	-	2	-	-
	Conv6	3x3	1024 → 512	1	Y	Y
	Sum (Conv6 + TransConv_Block4 output)	-	-	-	-	-
	BilinearUpSample2	-	-	2	-	-
	Conv7	3x3	512 → 256	1	Y	Y
	Sum (Conv7 + TransConv_Block3 output)	-	-	-	-	-
	BilinearUpSample3	-	-	2	-	-
	Conv8	3x3	256 → 128	1	Y	Y
	BilinearUpSample4	-	-	2	-	-
	Conv9	3x3	128 → 64	1	Y	Y
	Conv10	3x3	64 → 3	1	-	-

Table 7. The details of network parameters. TransConv_Block3-5 represent the three transposed convolution blocks in our framework (The diagrams can be found in Figure 2 in the main paper). SyncBatchNorm column indicates Synchronized Batch Normalization layer after Conv. ReLU column shows whether ReLU is used (following the SyncBatchNorm if SyncBatchNorm is used). BilinearUpSample represents bilinear upsampling. Sum denotes the simple summation. orig_H and orig_W are input image's height and width.

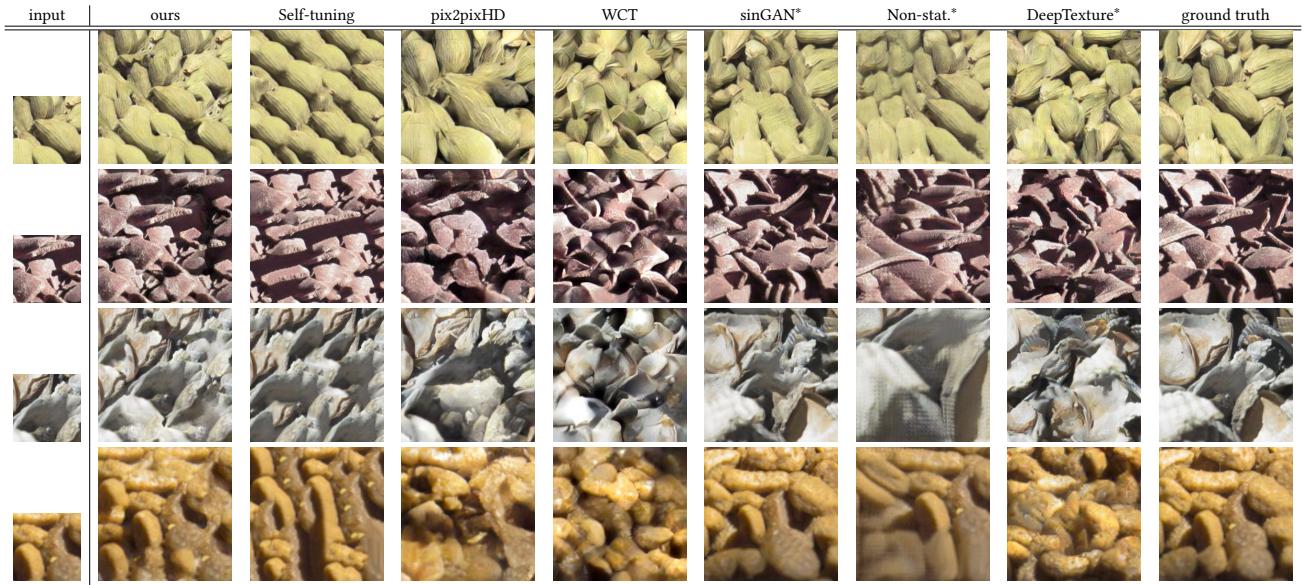


Fig. 13. Results of different approaches on 128 to 256 texture synthesis. sinGAN* Non-stat.* show the results of training with directly seeing the ground truth at target size. (Training with ground truth means using the ground truth 256×256 image as the target for each training iteration.) WCT is the style transfer based method. DeepTexture directly takes ground truth images as inputs.

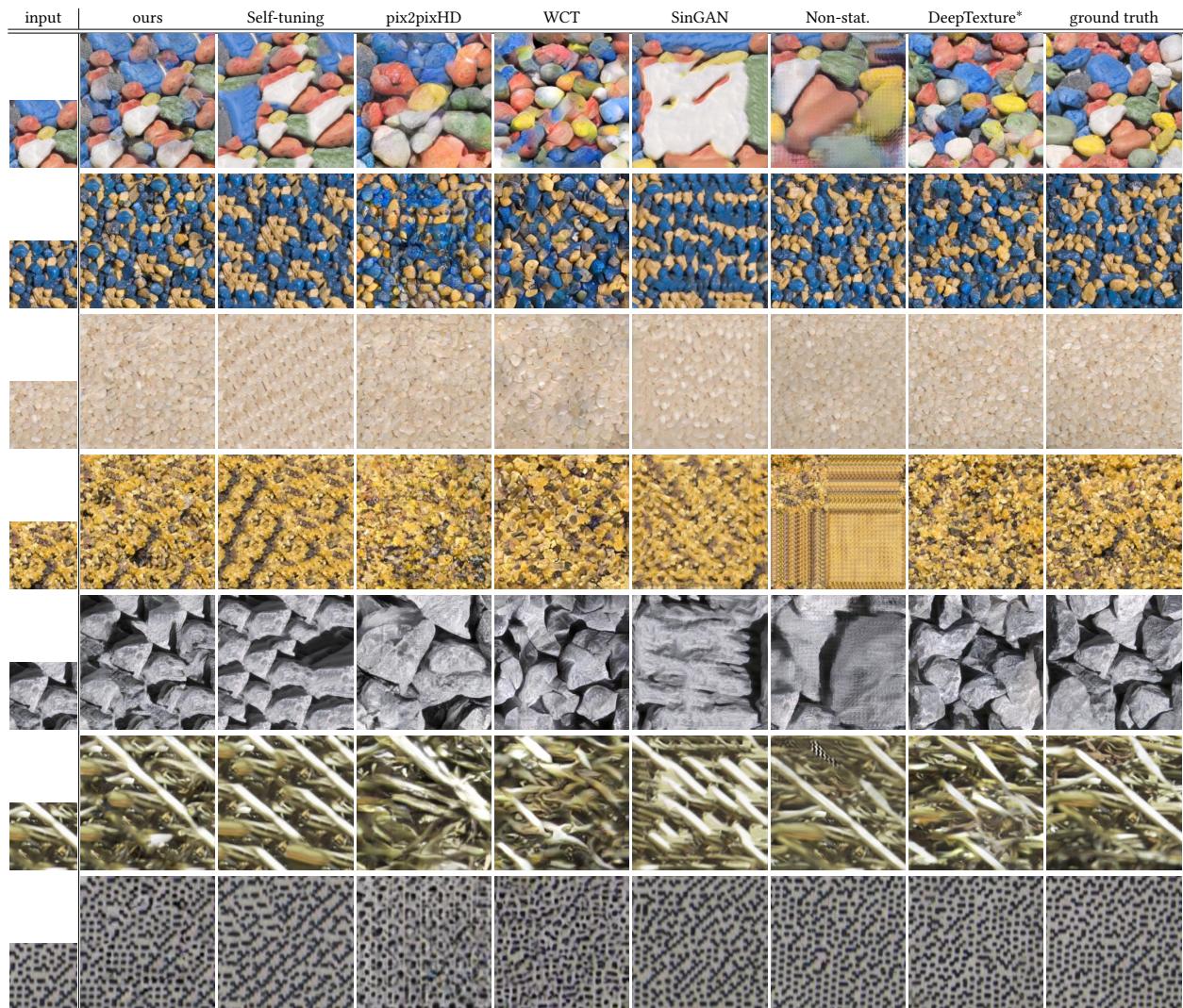


Fig. 14. Results of different approaches on 128 to 256 texture synthesis. SinGAN and Non-stat. results show the results of training without directly seeing ground truth at the exact target size.