

---

# The Unreasonable Effectiveness of Texture Transfer for Single Image Super-resolution

---

**Muhammad Waleed Gondal**

Max Planck Institute for Intelligent Systems.  
waleed.gondal@tuebingen.mpg.de

**Bernhard Schölkopf**

Max Planck Institute for Intelligent Systems.  
bs@tuebingen.mpg.de

**Michael Hirsch**

Amazon Research.  
hirsch@amazon.com

## Abstract

While implicit generative models such as GANs have shown impressive results in high quality image reconstruction and manipulation using a combination of various losses, we consider a simpler approach leading to surprisingly strong results. We show that texture loss[1] alone allows the generation of perceptually high quality images. We provide a better understanding of texture constraining mechanism and develop a novel semantically guided texture constraining method for further improvement. Using a recently developed perceptual metric employing “deep features” and termed LPIPS [2], the method obtains state-of-the-art results. Moreover, we show that a texture representation of those deep features better capture the perceptual quality of an image than the original deep features. Using texture information, off-the-shelf deep classification networks (without training) perform as well as the best performing (tuned and calibrated) LPIPS metrics. The code is publicly available.

## 1 Introduction

Recently, the task of single image super-resolution (SISR) has taken an interesting turn. Convolutional neural networks (CNNs) based models have not only been shown to reduce the distortions on full reference (FR) metrics for, e.g., PSNR, SSIM and IFC [3–8], but also to produce perceptually better images [4, 9]. The models trained specifically to reduce distortions fail at producing visually compelling results. They suffer from the issue of “regression-to-the-mean” as they mainly rely on minimizing the mean square error (MSE) between a high resolution image  $I_{HR}$  and an estimated image  $I_{est}$ , approximated from its low resolution counterpart  $I_{LR}$ . This minimization of MSE leads to the suppression of high frequency details in  $I_{est}$ , entailing blurred and over-smoothed images. Therefore, FR metrics do not conform with the human perception of visual quality. This was illustrated in [10, 11] and recently mathematically analyzed in [12].

The recently proposed methods [4, 9, 13] made substantial progress in improving the perceptual quality of the images by building on generative adversarial networks (GANs) [14]. The adversarial setting of a generator and a discriminator network helps the generator in hallucinating high frequency textures into the resultant images. Since the goal of the generator is to fool the discriminator, it may hallucinate fake textures which are not entirely faithful to the input image. This fake texture generation can be clearly observed in an  $8\times$  image super-resolution task. This behavior of GANs can be reduced using a combination of content preserving losses. This not only limits the ability of the generator to induce high quality textures but also makes it fall short in reproducing image details in the regions which have complex and irregular patterns such as tree leaves, rocks etc.

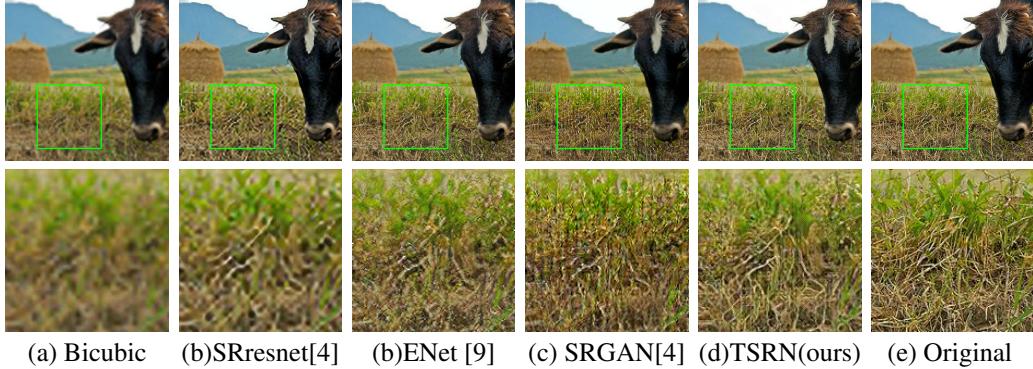


Figure 1: Visual Comparison of the recent state-of-the-art methods as measured by distortion and perceptual quality metrics with our texture based super-resolution network (TSRN) for  $4 \times$  SISR.

In the present paper we show that, in the task of SISR, perceptually high quality textures can be synthesized on the estimated images  $I_{est}$  using the Gram matrices based texture loss [1]. The loss was first employed by Gatys et al. in transferring realistic textures from a style image ( $I_s$ ) to a content image ( $I_c$ ). Despite the success of this method, the utility of texture transfer for enhancing natural images has not been studied extensively. This is because of the fact that while preserving the local spatial information of the textures, the texture loss discards the global spatial arrangement of the content image, rendering the semantic guidance of texture transfer a difficult problem.

We explore the effectiveness of Gram matrices in transferring and hallucinating realistic texture in the task of SISR. We show that despite its simplicity through the use of a single loss function, our proposed network yields favorable results when compared to state-of-the-art models that employ a mixture of loss functions and involve GANs that are notoriously difficult to train. In contrast, our model converges without the need of hand-tuned training schemes. We further build on this finding by providing external semantic guidance to control the texture transfer. We show that this scheme prevents the random spread of small features across object boundaries thus improving the visual quality of results especially in the challenging task of  $8 \times$  SISR. Furthermore, we demonstrate, that Gram matrices of deep features perform surprisingly well in measuring human perceived similarity between image patches.

## 2 Related Work

### 2.1 Super Resolution

Single image super-resolution(SISR) is the problem of approximating a high resolution ( $I_{HR}$ ) image from its corresponding low resolution ( $I_{LR}$ ) input image. The task is to fill in missing information in  $I_{HR}$  which involves the reconstruction and hallucination of textures, edges and low-level image statistics while remaining faithful to the low-resolution  $I_{LR}$  input. It is an under-determined inverse problem where different image priors have been explored to guide the upsampling of  $I_{LR}$  [15–17]. However, much success has been achieved by using recent data-driven approaches where a large number of training examples are used to set the prior over the empirical distribution of data. One of the earliest methods involved simple interpolation schemes [18], e.g. bicubic, Lanczos. Due to their simplicity and fast inference, these methods have been widely used, however they suffer from blurriness and can not predict high frequency details.

In recent years, learning based methods that try to learn a mapping between  $I_{LR}$  to  $I_{HR}$  have enjoyed much attention and progress. These methods can be classified into parametric and non-parametric methods [19]. Non-parametric algorithms include neighborhood embedding algorithms [20–23], that seek for the nearest match in an available database and try to synthesize an image by simple blending of different patches. Prone to mismatch and misalignment in patches these methods suffer from rendering artifacts in the HR output [24]. Parametric methods include sparse models [17], regression functions [8] and convolutional neural networks (CNNs). Dong et al. [7] first employed a shallow CNN to perform SISR on a bicubic interpolated image and got impressive results, [25]

successfully used a deep residual network. These CNN based methods use mean square error (MSE) as an optimization objective which leads to blurriness and fails to reconstruct high frequency details. Methods like [3, 4] tend to overcome this issue by minimizing perceptual losses in feature space. Ledig et al. [4] proposed SRResNet to show improvements in full-reference (FR) metrics. Follow-up work used a multi-scale optimized SRResNet architecture to win the NTIRE 2017 Super-Resolution Challenge [26] for 4x super-resolution. Moreover, [6] uses a coarse-to-fine laplacian pyramid framework to achieve state-of-the-art results in 8x super-resolution with respect to FR metrics.

More recently, GANs based methods [4, 9, 13] showed promising results by drastically improving the perceptual quality of images. In addition to the perceptual and adversarial losses used by [4], the patch-wise texture loss used by [9] helps synthesizing high quality textures. Our approach is different from [9], as we give up on the adversarial and perceptual loss terms. Moreover, we also don't use patch-wise texture loss and show that a globally applied texture loss is enough for spatially aligning textures and generating photo-realistic high-quality images. [27] also used patches and manually derived segmentation masks to constrain the texture synthesis in  $I_{est}$ . However, it highly relies on the efficiency of a slow patch matching algorithm and thus is prone to wrong matching of regions in  $I_{est}$  and  $I_{HR}$  which renders artifacts. The loss is also shown to be an important ingredient of a recent image-inpainting method [28]. A new deep features based contextual loss [29] is used by [30] to maintain the natural image statistics of  $I_{est}$ .

## 2.2 Neural Texture Transfer

The concept of neural texture transfer was first coined by Gatys et al. [1]. The method relies on matching the Gram matrices of VGG-19 [31] features to transfer the texture of one image to another. Afterwards, much work has been done in order to improve the speed [3, 32] and quality [33–35] of style transfer using feed forward networks and perceptual losses. Building on fast style transfer, [36, 37] proposed models to transfer textures from multiple style images. [34] showed improvement in style transfer by computing cross-layer Gram matrices instead of within-layer Gram matrices. Recently, Li et al. [38] has shown that matching the Gram matrices for style transfer is equivalent to minimizing MMD with the second order polynomial kernel. In addition to improving the style transfer mechanism, some work has been done to spatially constrain the texture transfer in order to maintain the textural integrity of different regions[39, 40]. Gatys et al. [40] demonstrated the spatial control of texture transfer using guided Gram matrices where binary masks are used as guidance channels in order to constrain the textures. Similar scheme was used by [33] in constraining style transfer. Instead of enforcing spatial guidance in the feature space of deep networks like these methods, we enforce it in pixel-space via customized texture loss which, unlike other methods, not only enables it to easily scale to multiple style images but also does not require semantic details at the test time.

Our main contributions are as follows:

- We provide a better understanding of texture constraining mechanism via texture loss and show that SISR of high perceptual quality can be achieved by using this as an objective function. The results compare well with GANs based methods on 4x SISR and outperform them on 8x SISR.
- Unlike GANs based methods, our method is easily reproducible and generates faithful textures especially in the constrained domain of facial images.
- To further enhance the quality of 8x SISR results, we formulate a novel semantically guided texture transfer scheme in order to avoid the intermixing of interclass textures such as grass, sky etc. The method is easily scalable to multiple style images and does not require semantic details at test time.
- We also show that Gram matrices provide a better and richer framework to capture the perceptual quality of images. Using this, our off-the-shelf deep classification networks (without training) perform as well as the best performing (tuned and calibrated) LPIPS metrics [2].

## 3 Texture Loss

The texture transfer loss was first proposed in the context of neural style transfer [1], where both style  $I_s$  and content images  $I_c$  are mapped into feature space using a VGG-19 architecture [31], pre-trained for image classification on image-net. The feature maps of both  $I_s$  and  $I_c$  are denoted by  $F^l \in \mathbb{R}^{N_l \times M_l}$  and  $P^l \in \mathbb{R}^{N_l \times M_l}$  respectively, where  $N_l$  is the number of feature maps in layer  $l$

and  $M_l$  is the product of height and width of feature maps in layer  $l$  i.e.  $M_l = \text{height} \times \text{width}$ . A Gram matrix is the inner product of vectorized feature maps. Therefore the Gram matrices for both  $F^l$  and  $P^l$  are computed as  $G_{i,j}^l = \mathbf{F}_i^T \mathbf{F}_j$  and  $A_{i,j}^l = \mathbf{P}_i^T \mathbf{P}_j$ . The texture loss  $\mathcal{L}_{\text{texture}}$  is defined by the mean squared error between the feature correlations expressed by these Gram matrices.

$$\mathcal{L}_{\text{texture}} = \frac{1}{4N_l^2 M_l^2} \sum_{i=1}^{N_l} \sum_{j=1}^{M_l} (G_{i,j}^l - A_{i,j}^l)^2 \quad (1)$$

The loss tries to match the global statistics of  $I_c$  with  $I_s$ , captured by the correlations between feature responses in layers  $l$  of the VGG-19. These correlations capture the local spatial information in the feature maps while discard their global spatial arrangement [41].

### 3.1 Constraining Texture Transfer

The above loss tries to match the global level statistics of  $I_s$  and  $I_c$  without retaining the spatial arrangement of the content image. However, we observe that if there exists a good feature space correspondence between  $I_s$  and  $I_c$  then the Gram matrices alone constrain the texture transfer such that it preserves the semantic details of the content image. The composition of Gram matrices makes use of the translational invariance property of the pre-trained VGG-19's [31] convolutional kernels in mapping the textures correctly. We shed more light on this texture constraining mechanism and its translational invariant mapping in the appendix. Thus Gram matrices' provide a stable spatial control such that the texture from  $I_s$  maps to the corresponding features on  $I_c$ . Fig 2 shows texture transfer of a non-texture image for different initial approximates of  $I_c$  using iterative optimization approach by [1]. Second column depicts the results of vanilla style transfer [1] on a plain white image, 4x upsampled image and an 8x upsampled images respectively. In case of plain white image, the texture gets transferred in an uncontrollable fashion. This is the known phenomenon in image style transfer. However, the texture transfer on a 4x and 8x upsampled images shows consistency in texture mapping i.e. texture from  $I_s$  gets mapped to the correct corresponding regions of  $I_c$ . We observe that the interpolated approximates  $I_{\text{est}}$  of  $I_{LR}$  are good enough for establishing feature-space correspondences and thus mapping the textures correctly.

In the Fig 2, one can observe that the texture transfer for a 4x interpolated image is much better than that for an 8x. The ambiguousness in texture transfer for an 8x upsampled  $I_{LR}$  is because of the absence of enough content features to establish correspondences. Thus to better guide the texture transfer in 8x SISR, we devise an external semantic guidance scheme. The third column in Fig 2 shows the effectiveness of the semantically guided texture transfer. In comparison to the second column we can see that the texture is transferred in a more coherent fashion.

### 3.2 Texture Loss in SISR.

In SISR we try to find a mapping between a low-resolution input image  $I_{LR}$  and a high-resolution output image  $I_{HR}$ . As a function approximator we use a deep CNN. While recent state-of-the-art methods use a combination of various loss functions, our texture super resolution network (TSRN) is specifically trained to optimize for  $\mathcal{L}_{\text{texture}}$  in equation 1 which yields images of perceptually high quality for 4 $\times$  and 8 $\times$  super-resolution, Fig 5 and 6.

### 3.3 SISR via Semantically Constrained Textures.

In order to make full use of the texture loss based image super resolution, we also performed externally controlled semantic texture transfer. We enforce semantic details via loss function. For the implementation of semantic control of texture transfer, we use the ground truth segmentation masks provided by the recently released dataset MS-COCO stuff dataset [42].

Additional spatial control is provided by making use of the semantic information present inside an image. Instead of matching the global level statistics of an image we divide the image into  $r$  segments semantically. Each segment exhibits its own local level statistics which are different from the other segments of the same image. This facilitates us to match the local level statistics at an individual segment level. Also it helps in preserving the global spatial arrangement of the segments as the relative spatial information of each segment is considered before extracting them from the images.



(a) Texture/Content    (b) no semantic guidance [1]    (c) with semantic guidance

Figure 2: (a) shows  $I_{HR}$  (in insets) and a plain white, 4x and 8x upsampled versions of  $I_{HR}$  as  $I_c$ . (b) vanilla neural texture transfer [1]. (c) neural texture transfer with semantic guidance.

Our method gains inspiration from the spatial control of texture transfer based on guided Gram matrices (GGMs) [40] where binary segmentation masks are used to define which region of a style image would get mapped to the specific region of a content image. It uses  $r$  segmentation masks  $I_{seg}^r$  to compute guidance channels ( $\mathbf{T}_l^r$ ) for each layer  $l$  of a CNN by either down-sampling them to match the dimensions of each layer’s feature maps or by enforcing spatial guidance only on neurons whose receptive field lie inside the guidance region for better results. The guidance channels are then used to form spatially guided feature maps by the element-wise multiplication of texture image features and the guidance channels. This method of computing GGMs for training a deep architecture is not feasible, especially in our case where we have multiple segmentation masks for each image. As the computation of spatially guided feature maps depends on the element-wise multiplication of ( $\mathbf{T}_l^r$ ) with texture image feature maps which depending on the size and depth of chosen CNN layers can be computationally very expensive. Moreover, the computation of better guidance channels require extra efforts of discarding the neurons whose receptive fields lie outside of guidance region. We propose a simplification of this process by removing the need of guidance channels ( $\mathbf{T}_l^r$ ) and the explicit computation of spatially guided feature maps altogether. The  $r$  binary segmentation masks  $I_{seg}^r$  (having pixel value of 1 for the class of interest and 0 elsewhere) where each mask categorically represents a different region of an image are element-wise multiplied with the texture image  $I_{HR}$  and the estimated image  $I_{est}$  to give out  $I_{target}^r$  and  $I_{est}^r$  respectively, Fig 3.

$$I_{target}^r = I_{HR} \circ I_{seg}^r \quad (2)$$

$$I_{est}^r = I_{est} \circ I_{seg}^r \quad (3)$$

These segmented images are then propagated to the VGG19 and Gram matrices of their feature maps are then computed in normal fashion. The method is flexible and relatively fast to enforce spatial guidance of texture transfer, especially when it has to be used for training a deep architecture. The texture loss is then performed individually for all the segmented images. Equation 4 shows the

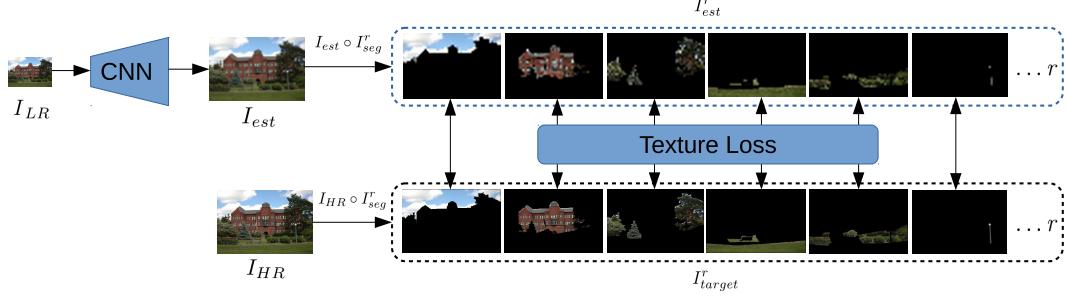


Figure 3: Scheme for semantically controlled texture transfer.

objective function formulation of the complete semantically controlled texture transfer. See abstract to check the effectiveness of our proposed semantically controlled fast style transfer.

$$\mathcal{L}_{texture} = \sum_{k=1}^r \frac{1}{4N_l^2 M_l^2} \sum_{i=1}^{N_l} \sum_{j=1}^{M_l} (G_{i,j}^l(I_{target}^k) - A_{i,j}^l(I_{est}^k))^2 \quad (4)$$

## 4 Architecture

For the implementation of TSRN, we employ a fully convolutional neural network architecture inspired by [9]. The architecture is efficient at inference time as it performs most feed forward computations on  $I_{LR}$  and is deep enough to perform texture synthesis. The presence of residual blocks facilitates convergence during training. Similarly to [9], we also add a bi-cubically upsampled verison of  $I_{LR}$  to the predicted output such that the network is only required to learn the residual image. This helps to reduce color shifts during training as also reported by [9]. However, instead of using nearest neighbor up-sampling, we use a pixel resampling layer [43] because of its recent proven success in generative networks [44]. The method is also shown to be agnostic to model's depth. See appendix for more details.

## 5 Implementation

We trained our network on MS-COCO [42], where we center crop image patches sized  $256 \times 256$  pixels. The patches are then bi-cubically down-sampled  $4 \times$  or  $8 \times$  to  $64 \times 64$  or  $32 \times 32$ , respectively. We first pretrain our network by minimizing mean square error (MSE) for 10 epochs. We found this pre-training beneficial for the subsequent Gram matrix based optimization as it facilitates the detection of relevant features for texture transfer. After pretraining, we train our model using only 1

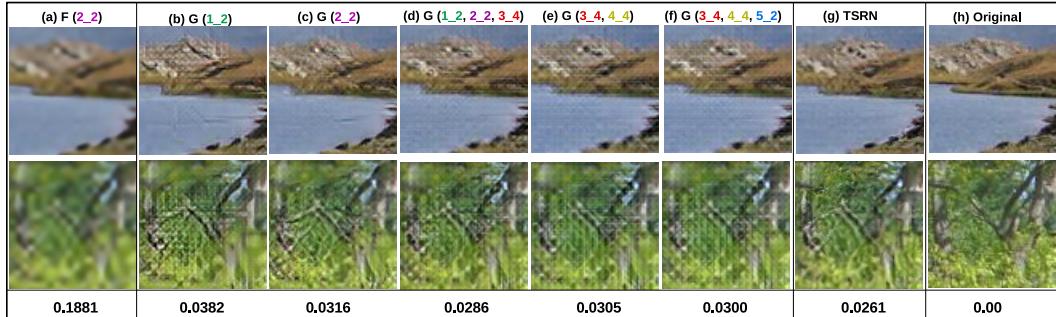


Figure 4: Layer and loss ablation study on SunHays dataset [24]. Each column shows the effects of different VGG19[31] layers on the visual quality of a restored image. Perceptual loss using deep features (F) generates blurred images (left most column) in comparison to Gram matrices (G) based restoration. The last row shows the mean LPIPS score on the dataset (lower score is better).

as an objective function for another 100 epochs. We found that the network converges after approximately 60 epochs. For the implementation of  $\mathcal{L}_{texture}$ , we compute Gram matrices on layers *conv2\_2*, *conv3\_4*, *conv4\_4* and *conv5\_2* of a pre-trained VGG-19 architecture. To justify the selection of specific VGG-19’s layers for texture loss, we provide a qualitative and quantitative (LPIPS) analysis on SunHays dataset in Figure4. We considered convolutional layers before each pooling layer except (*conv1\_2*) as this layer, containing more pixel-level and less structural information, causes artifacts and over-smoothing in images. The selection of only higher layers tend to generate checkboard artifacts. In Figure4, all the networks are trained using the same architecture and procedure mentioned in the paper for 100 epochs. The network is trained with the learning rate of 0.0005 using ADAM as an optimizer. We use the PyTorch framework [45] to implement the model on a Nvidia Tesla P40 GPU. Inference time for 4 $\times$  and 8 $\times$  SISR is approximately 41 and 32 milliseconds for a 1 mega-pixel image and 0.203 and 0.158 seconds for a 5 mega-pixel image on the GPU.

For our results on segmentation based super-resolution (TSRN-S), we pre-train on the MS-COCO dataset before we train on the MS-COCO stuff dataset using equation 4 as an objective function. The stuff dataset is particularly suited for our task as it not only contains the segmentation masks of object instances but also outdoor scenes like grass, sky, buildings etc. Statistically, these regions cover more than 60% [46] of images showing natural scenes. To reduce the computation time, we consider the binary segmentation masks of only six maximally represented classes in each image (based on their pixel count). Whereas the seventh mask covers the ‘others’ class, containing the remaining regions of the image. If there are less than six classes in an image then the ‘others’ class is replicated to give out seven masks per image. Both code and pre-trained models are publicly available.

## 6 Experimental Results

We evaluate both our proposed models, one with globally computed Gram matrices (TSRN-G) and semantically guided Gram matrices (TSRN-S). We provide extensive quantitative and qualitative comparisons with recent state-of-the-art methods in SISR as measured by distortion metrics (PSNR, SSIM etc.) but also in terms of perceptual quality.

### 6.1 Quantitative Evaluation

Distortion metrics have been shown to not correlate well with perceived image quality [9, 2]. Therefore, for quantitative comparison we follow [9] and report the performance in object recognition as a proxy for perceived image quality. Additionally, we report numbers for a recently proposed learned full-reference image quality metric [2] that approximates perceptual similarity.

#### 6.1.1 Object Recognition Performance.

The perceptual quality of an image correlates very well with its performance on object recognition models which are trained on the large corpus of image-net, as corroborated by [9]. Recently, the same methodology of assessing image quality has been adopted by a competition<sup>1</sup>. The idea stems from the observation that deep features are very effective in capturing the perceptual quality of images [2]. Therefore, we perform our comparison with other methods utilizing the standard image classification models trained on ImageNet. We randomly pick 1000 images from the ILSVRC 12 validation

TopK	Methods	Bicubic	SRResNet[4]	SRGAN[4]	ENet-PAT[9]	TSRN-S	TSRN-G	Baseline
Top 1	DenseNet-169	0.594	0.641	0.666	0.658	0.688	<b>0.692</b>	0.713
	ResNet-50	0.545	0.616	0.655	0.649	<b>0.674</b>	0.671	0.703
	VGG-19	0.455	0.538	0.578	0.571	<b>0.610</b>	0.609	0.656
Top 5	DenseNet-169	0.788	0.862	0.864	0.857	<b>0.876</b>	0.871	0.890
	ResNet-50	0.776	0.841	0.847	0.843	0.862	<b>0.866</b>	0.885
	VGG-19	0.676	0.772	0.798	0.792	0.819	<b>0.821</b>	0.853

Table 1: Top-1 and Top-5 image recognition accuracy on 4 $\times$  SISR images.

<sup>1</sup><http://www.ug2challenge.org/>

TopK	Methods	Bicubic	SRResNet[4]	SRGAN[4]	TSRN-S	TSRN-G	Baseline
Top 1	DenseNet-169	0.353	0.506	0.432	<b>0.509</b>	0.506	0.713
	ResNet-50	0.301	0.437	0.424	0.484	<b>0.503</b>	0.703
	VGG-19	0.239	0.343	0.267	0.374	<b>0.389</b>	0.656
Top 5	DenseNet-169	0.602	0.727	0.676	0.733	<b>0.743</b>	0.890
	ResNet-50	0.518	0.689	0.657	<b>0.718</b>	0.717	0.885
	VGG-19	0.406	0.565	0.504	<b>0.613</b>	0.611	0.853

Table 2: Top-1 and Top-5 image recognition accuracy on  $8 \times$  SISR images.

Metric	Set 5		Set 14		BSD 100		Urban	
	AlexNet	SNet	AlexNet	SNet	AlexNet	SNet	AlexNet	SNet
Bicubic	0.1585	0.1202	0.1731	0.1320	0.1463	0.1007	0.1552	0.1238
SRCNN[7]	0.0964	0.0732	0.1175	0.1025	0.1257	0.0920	0.0960	0.0905
LapSRN[6]	0.0566	0.0556	0.1002	0.0967	0.1005	0.0753	0.0746	0.0757
MSLapSRN[47]	0.0551	0.0574	0.0972	0.0916	0.0989	0.0720	0.0691	0.0709
SRResNet[4]	0.0538	0.0491	0.0848	0.0821	0.0909	0.0625	0.0628	0.0652
SRGAN[4]	0.0275	0.0466	0.0575	0.0679	0.0484	0.0527	0.0401	0.0584
ENet-PAT[9]	<b>0.0251</b>	0.0391	0.0569	0.0590	0.0494	0.0472	0.0414	0.0467
TSRN-S (Ours)	0.0273	0.0394	<b>0.0438</b>	0.0483	<b>0.0478</b>	0.0420	0.0397	0.0404
TSRN-G (Ours)	0.0285	<b>0.0358</b>	0.0463	<b>0.0456</b>	0.0481	<b>0.0404</b>	<b>0.0385</b>	<b>0.0392</b>

Table 3: Comparison for  $4 \times$  SISR on pre-trained AlexNet-linear and SqueezeNet-linear LPIPS metric[2]. Lower score is better.

dataset and super-resolve their downsampled versions using different super-resolution models. The performance is evaluated on how much recognition accuracy is retained by each model, compared to the baseline accuracy. Tables 1 and 2 show that our proposed TSRN model outperforms all other state-of-the-art SISR methods for both  $4 \times$  and  $8 \times$  super-resolution.

### 6.1.2 LPIPS.

The Learned Perceptual Image Patch Similarity (LPIPS) metric [2] is a recently introduced full-reference image quality assessment metric which tries to measure the perceptual similarity between two images. The metric uses linearly calibrated off-the-shelf standard deep classification networks trained to measure the perceptual similarity of the images. The networks are trained on the very large Berkeley-Adobe Perceptual Patch Similarity (BAPPS) [2] dataset, containing human perceptual judgments. We use the pre-trained, linearly calibrated AlexNet and SqueezeNet networks<sup>2</sup>. The networks are trained on patches sized  $64 \times 64$  pixels. Therefore, we also divide the images into patches of size  $64 \times 64$  pixels. For each image, we pick its shorter dimension and find the nearest possible value  $v$  divisible by 64, then we center crop an image of resolution  $v \times v$ . The cropped image is then further divided into patches of size  $64 \times 64$ . We report the averaged perceptual similarity determined on those patches.

In Table 3 we use the recommended AlexNet (linear) and SqueezeNet (linear) models for measuring the perceptual quality. We found the quantitative evaluations to be consistent across numerous models that have been trained to improve either PSNR, SSIM scores such as SRResNet, LapSRN, SRCNN or the ones trained to improve perceptual quality such as SRGAN and ENet-PAT. TSRN consistently achieves better perceptual similarity scores than other methods.

## 6.2 Visual Comparison

In Fig. 5 and 6 we show visual comparisons with recently proposed state-of-the-art models for both  $4 \times$  and  $8 \times$  super-resolution. Our TSRN model manages to hallucinate realistic textures and image details and compares favorably with the state-of-the-art.

<sup>2</sup><https://github.com/richzhang/PerceptualSimilarity>

Metric	Set 5		Set 14		BSD 100		Urban	
	AlexNet	SNet	AlexNet	SNet	AlexNet	SNet	AlexNet	SNet
Bicubic	0.27464	0.22877	0.27390	0.24669	0.22802	0.20202	0.23854	0.22946
LapSRN[6]	0.19849	0.15506	0.21525	0.19058	0.19009	0.16379	0.15638	0.15426
MSLapSRN[47]	0.16748	0.13609	0.20184	0.17599	0.17679	0.15276	0.13252	0.13328
SRResNet[4]	0.13679	0.11958	0.18091	0.16060	0.16148	0.13512	0.13714	0.13217
SRGAN[4]	0.14230	0.15007	0.13801	0.12720	0.13276	0.10902	0.12929	0.12470
TSRN-S(Ours)	<b>0.0859</b>	0.0863	<b>0.1194</b>	<b>0.0963</b>	<b>0.1021</b>	<b>0.0823</b>	0.0918	<b>0.0802</b>
TSRN-G(Ours)	0.0900	<b>0.0859</b>	0.1277	0.1092	0.1029	0.0833	<b>0.0900</b>	0.0817

Table 4: Comparison for  $8\times$  SISR on pre-trained AlexNet-linear and SqueezeNet-linear LPIPS Perceptual Similarity Metric models. Lower score is better.

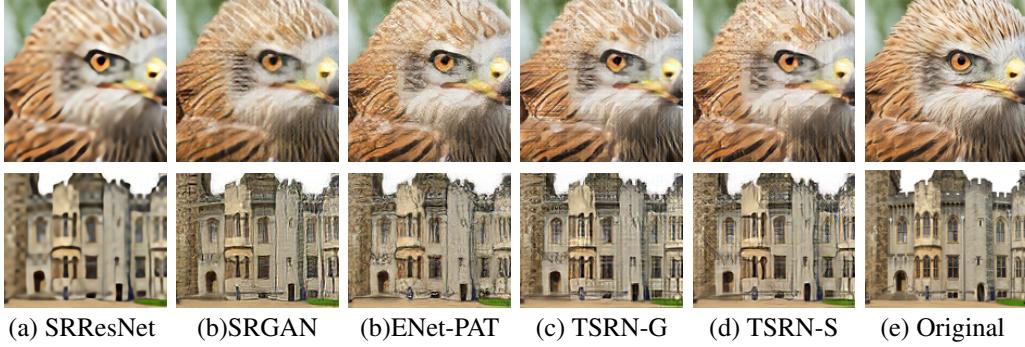


Figure 5: Visual Comparison of recent state-of-the-art methods based on distortion metrics and perceptual quality with our texture based  $4\times$  image super-resolution.

### 6.3 TSRN-Faces on CelebA Dataset.

In addition to training on MS-COCO dataset[42], we also tested our proposed texture based super resolution method for CelebA faces dataset[48]. Our method yields visible improvements over other methods. More specifically we compare with Enhancenet-PAT [9] which employs GAN for enhancing textures. We observe that such method has a tendency to manipulate the overall facial features, thus not maintaining the integrity of the input image. In comparison, our method learns the texture mapping between a low resolution image ( $I_{LR}$ ) and its high resolution counterpart ( $I_{HR}$ ) thus generates visually plausible results.

## 7 Using Texture as a Perceptual Metric

In this section, we propose an improvement on LPIPS [2], a recently proposed perceptual similarity metric based on deep features. The method computes the distance between the deep features of two images in order to determine the perceptual similarity between them. We argue that Gram matrices that measure the correlations of the same deep features, provide a richer and better framework for capturing the perceptual representation of images than the features themselves. Therefore, instead of computing the distances between the features of a given convolutional layer, we compute the distance between their Gram matrices. For a pair of reference and distorted patches ( $x, x_0$ ), we compute their normalized Gram matrices  $\hat{G}^l$  and  $\hat{A}^l \in \mathbb{R}^{C_l \times C_l}$ , where  $C$  is the number of channels in layer  $l$ . We compute the distance between them using the same formulation as in equation1 and then sum it up across all layers  $l$ , i.e.

$$d(x, x_0) = \sum_l \frac{1}{C_l^2} \sum_{i=1}^{C_l} \sum_{j=1}^{C_l} (G_{i,j}^l - A_{i,j}^l)^2 \quad (5)$$

Using the features of “uncalibrated” pre-trained image classification networks, this Gram matrices distance achieves better 2AFC scores on the BAPPS validation dataset than the distances based

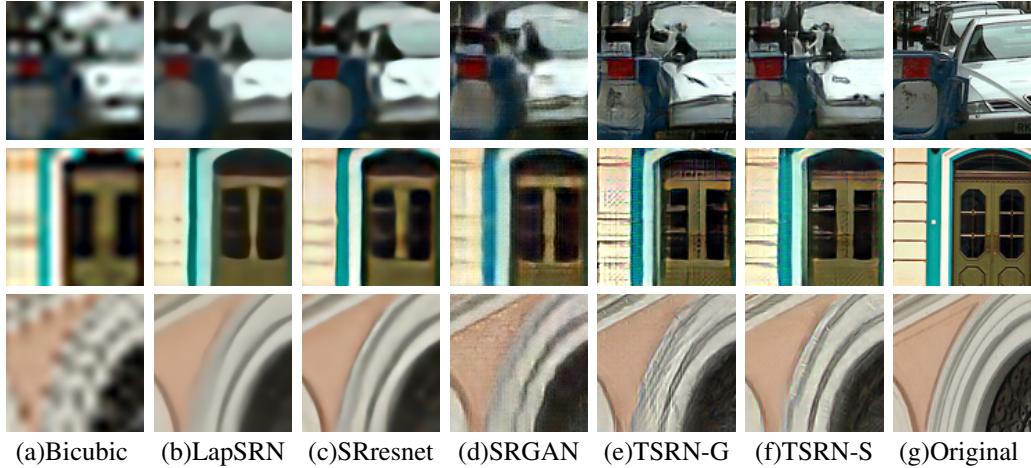


Figure 6: Visual Comparison of recent state-of-the-art methods based on distortion metrics and perceptual quality with our texture based  $8\times$  image super-resolution.

on the features themselves. As shown in Figure 8, our results (Net-G) are comparable to the ‘‘calibrated’’ LPIPS models (specifically trained on BAPPS training datasets) and also outperform them in some benchmarks. For comparison, we adopted the same configuration of three reference models (SqueezeNet[49], AlexNet[50] and VGG-16[31]) used by [2]. However, to get the best results we changed the number of layers for the distance computation, more specifically we did not use the feature activations before the first pooling layer and after the penultimate pooling layer of each model. This is because the texture from the lowest layers do not contain any structure in them whereas the last layers capture abstract and semantically more meaningful representation of the image but they lack in their ability to capture the perceptual details [41].

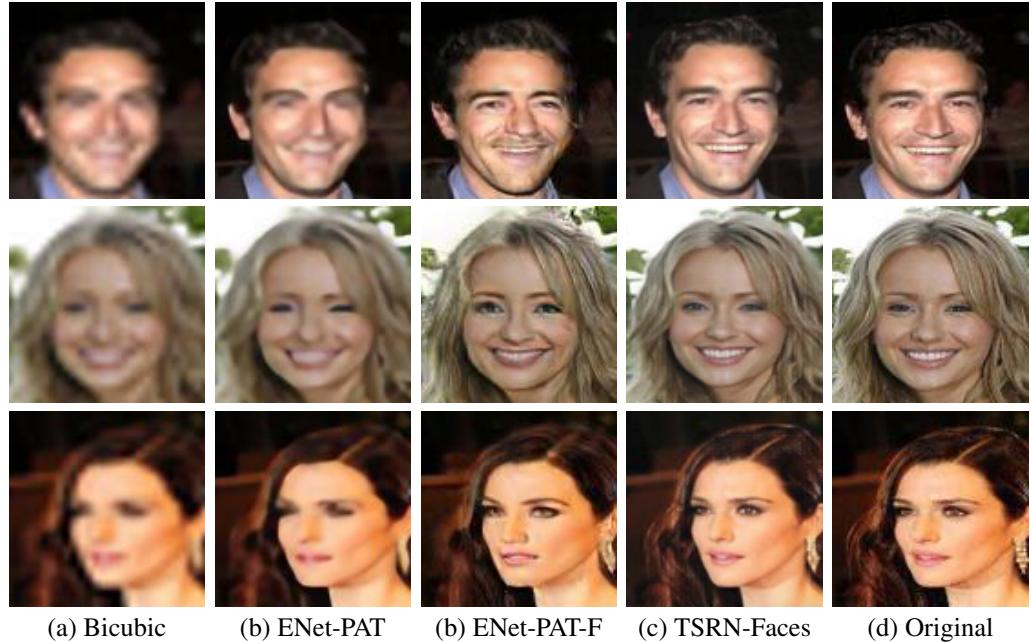


Figure 7: Visual comparison of different networks trained on CelebA dataset [48] for  $4\times$  SISR. TSRN yields visually faithful results to the original input image.

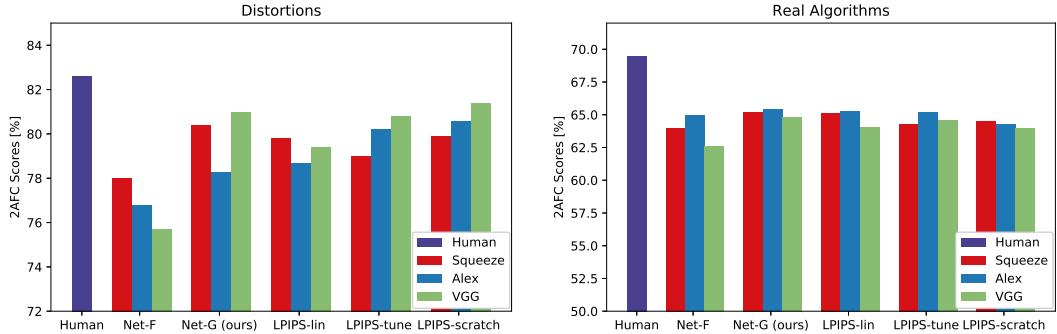


Figure 8: Quantitative comparison between different methods for determining perceptual similarity on the BAPPS validation dataset [2]. Our Gram matrices based distance (Net-G) scores better than the feature based method (Net-F). Net-G results are comparable to calibrated \*LPIPS metrics which are specifically trained on BAPPS training dataset, thus have an advantage.

Subtype	Metric	Distortions			Real Algorithms				All	
		Trad- itional	CNN- Based	All	Super- res	Video Deblur	Color- ization	Frame Interp	All	All
Oracle	Human	80.8	84.4	82.6	73.4	67.1	68.8	68.6	69.5	73.9
*LPIPS [2]	Squeeze – lin	76.1	<b>83.5</b>	79.8	71.1	60.8	<b>65.3</b>	63.2	65.1	<b>70.0</b>
	Alex – lin	73.9	83.4	78.7	71.5	<b>61.2</b>	<b>65.3</b>	63.2	<b>65.3</b>	69.8
	VGG – lin	76.0	82.8	79.4	70.5	60.5	62.5	63.0	64.1	69.2
	Squeeze – scratch	74.9	83.1	79.0	71.1	60.8	63.0	62.4	64.3	69.2
	Alex – scratch	77.6	82.8	80.2	71.1	61.0	<b>65.6</b>	<b>63.3</b>	65.2	<b>70.2</b>
	VGG – scratch	77.9	<b>83.7</b>	80.8	71.1	60.6	64.0	62.9	64.6	70.0
	Squeeze – tune	76.7	83.2	79.9	70.4	<b>61.1</b>	63.2	63.2	64.5	69.6
	Alex – tune	77.7	83.5	80.6	69.1	60.5	64.8	62.9	64.3	69.7
	VGG – tune	<b>79.3</b>	<b>83.5</b>	<b>81.4</b>	69.8	60.5	63.4	62.3	64.0	69.8
Supervised- Nets [2]	SqueezeNet [49]	73.3	82.6	78.0	70.1	60.1	63.6	62.0	64.0	68.6
	AlexNet [50]	70.6	83.1	76.8	<b>71.7</b>	60.7	65.0	62.7	65.0	68.9
	VGG [31]	70.1	81.3	75.7	69.0	59.0	60.2	62.1	62.6	67.0
Supervised- Nets (Ours)	SqueezeNet [49]	77.5	83.2	80.4	<b>71.6</b>	<b>61.1</b>	65.1	62.9	65.2	<b>70.2</b>
	AlexNet [50]	73.5	83.0	78.3	71.5	60.9	<b>65.6</b>	<b>63.4</b>	<b>65.4</b>	69.7
	VGG [31]	<b>78.3</b>	<b>83.7</b>	<b>81.0</b>	70.9	60.9	64.3	63.1	64.8	<b>70.2</b>

Table 5: 2AFC scores (higher is better) for different methods using disparity in deep feature representations [2] and texture representations (ours) on BAPPS validation dataset. Values in blue are **highest performing** while the values in red are **the second best**. Our texture based scores from untrained supervised networks consistently perform better than the feature based scores and compare to \*LPIPS metrics which are specifically trained on BAPPS training dataset, thus have an advantage over other untrained methods.

## 8 Conclusion

Transferring texture via matching Gram matrices has been very successful in image style transfer, however their utility for natural image enhancement has not been studied extensively. In this work we demonstrate that Gram matrices are very powerful in capturing perceptual representations of images which makes them a perfect candidate for their use in a perceptual similarity metric like LPIPS. Exploiting this ability, we obtain image reconstructions of high perceptual quality for the task of  $4\times$  and  $8\times$  single image super-resolution. We further devise a scheme for external semantic guidance for controlling texture transfer which is particularly helpful for  $8\times$  super-resolution. Our method is simple, easily reproducible and yet effective. We believe that texture loss can have far reaching implications in the future research of image restoration.

## References

- [1] Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2414–2423
- [2] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. arXiv preprint arXiv:1801.03924 (2018)
- [3] Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, Springer (2016) 694–711
- [4] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint arXiv:1609.04802 (2016)
- [5] Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (2017)
- [6] Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition. (2017)
- [7] Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European Conference on Computer Vision, Springer (2014) 184–199
- [8] Kim, K.I., Kwon, Y.: Single-image super-resolution using sparse regression and natural image prior. IEEE transactions on pattern analysis and machine intelligence **32**(6) (2010) 1127–1133
- [9] Sajjadi, M.S.M., Schölkopf, B., Hirsch, M.: EnhanceNet: Single image super-resolution through automated texture synthesis (2017)
- [10] Laparra, V., Ballé, J., Berardino, A., Simoncelli, E.P.: Perceptual image quality assessment using a normalized laplacian pyramid. Electronic Imaging **2016**(16) (2016) 1–6
- [11] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4) (2004) 600–612
- [12] Blau, Y., Michaeli, T.: The perception-distortion tradeoff. arXiv preprint arXiv:1711.06077 (2017)
- [13] Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. arXiv preprint arXiv:1804.02815 (2018)
- [14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
- [15] Zoran, D., Weiss, Y.: From learning models of natural image patches to whole image restoration. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 479–486
- [16] Sun, J., Xu, Z., Shum, H.Y.: Image super-resolution using gradient profile prior. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
- [17] Liu, D., Wang, Z., Wen, B., Yang, J., Han, W., Huang, T.S.: Robust single image super-resolution via deep networks with sparse prior. IEEE Transactions on Image Processing **25**(7) (2016) 3194–3207
- [18] Hou, H., Andrews, H.: Cubic splines for image interpolation and digital filtering. IEEE Transactions on acoustics, speech, and signal processing **26**(6) (1978) 508–517
- [19] Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 5197–5206
- [20] Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-based super-resolution. IEEE Computer graphics and Applications **22**(2) (2002) 56–65
- [21] Chang, H., Yeung, D.Y., Xiong, Y.: Super-resolution through neighbor embedding. In: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. Volume 1., IEEE (2004) I–I
- [22] Timofte, R., De Smet, V., Van Gool, L.: A+: Adjusted anchored neighborhood regression for fast super-resolution. In: Asian Conference on Computer Vision, Springer (2014) 111–126

- [23] Yang, J., Lin, Z., Cohen, S.: Fast image super-resolution based on in-place example regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 1059–1066
- [24] Sun, L., Hays, J.: Super-resolution from internet-scale scene matching. In: Computational Photography (ICCP), 2012 IEEE International Conference on, IEEE (2012) 1–12
- [25] Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1646–1654
- [26] Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L., Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M., et al.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, IEEE (2017) 1110–1121
- [27] Sun, L., Hays, J.: Super-resolution using constrained deep texture synthesis. arXiv preprint arXiv:1701.07604 (2017)
- [28] Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. arXiv preprint arXiv:1804.07723 (2018)
- [29] Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. arXiv preprint arXiv:1803.02077 (2018)
- [30] Mechrez, R., Talmi, I., Shama, F., Zelnik-Manor, L.: Learning to maintain natural image statistics, [arxiv](<https://arxiv.org/abs/1803.04626>). arXiv preprint arXiv:1803.04626 (2018)
- [31] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
- [32] Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.S.: Texture networks: Feed-forward synthesis of textures and stylized images. In: ICML. (2016) 1349–1357
- [33] Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. CoRR, abs/1703.07511 (2017)
- [34] Yeh, M.C., Tang, S.: Improved style transfer by respecting inter-layer correlations. arXiv preprint arXiv:1801.01933 (2018)
- [35] Liao, J., Yao, Y., Yuan, L., Hua, G., Kang, S.B.: Visual attribute transfer through deep image analogy. arXiv preprint arXiv:1705.01088 (2017)
- [36] Dumoulin, V., Shlens, J., Kudlur, M., Behboodi, A., Lemic, F., Wolisz, A., Molinaro, M., Hirche, C., Hayashi, M., Bagan, E., et al.: A learned representation for artistic style. arXiv preprint arXiv:1610.07629 (2016)
- [37] Chen, D., Yuan, L., Liao, J., Yu, N., Hua, G.: Stylebank: An explicit representation for neural image style transfer. arXiv preprint arXiv:1703.09210 (2017)
- [38] Li, Y., Wang, N., Liu, J., Hou, X.: Demystifying neural style transfer. arXiv preprint arXiv:1701.01036 (2017)
- [39] Lu, M., Zhao, H., Yao, A., Xu, F., Chen, Y., Zhang, L.: Decoder network over lightweight reconstructed feature for fast semantic style transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 2469–2477
- [40] Gatys, L.A., Ecker, A.S., Bethge, M., Hertzmann, A., Shechtman, E.: Controlling perceptual factors in neural style transfer. arXiv preprint arXiv:1611.07865 (2016)
- [41] Gatys, L., Ecker, A.S., Bethge, M.: Texture synthesis using convolutional neural networks. In: Advances in Neural Information Processing Systems. (2015) 262–270
- [42] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer (2014) 740–755
- [43] Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: 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) 1874–1883
- [44] Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)

- [45] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. (2017)
- [46] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. arXiv preprint arXiv:1608.05442 (2016)
- [47] Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Fast and accurate image super-resolution with deep laplacian pyramid networks. arXiv:1710.01992 (2017)
- [48] Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV). (2015)
- [49] Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016)
- [50] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
- [51] Dieleman, S., Willett, K.W., Dambre, J.: Rotation-invariant convolutional neural networks for galaxy morphology prediction. Monthly notices of the royal astronomical society **450**(2) (2015) 1441–1459

## 9 Appendix

### 9.1 Constraining Texture Transfer.

In this section we try to understand how different textures get hallucinated and how important is the role of content image's features in determining the mapping of textures from a style image.

It is necessary to understand the role of pre-trained VGG-19's [31] convolutional kernels which form the basis of Gram matrices composition. The kernels are translational invariant as they are able to look out for the specific features (for example, an oriented edge in an image) using the same weights after sweeping across the entire image. A specific feature in the entire image space would always be detected by a specific kernel [51]. This means that if both  $I_s$  and  $I_c$  have identical features then they are always going to be detected by the same kernels. Therefore, the activations in resulting feature maps  $F^l \in \mathbb{R}^{N_l \times M_l}$  and  $P^l \in \mathbb{R}^{N_l \times M_l}$ , where  $N_l$  is the number of feature maps in layer  $l$  and  $M_l$  is the product of height and width of feature maps in layer  $l$  i.e.  $M_l = \text{height} \times \text{width}$ , would correspond to each other in channel dimensions  $N_l$  (each kernel results in one feature map). However, the kernels are not rotational and scale invariant thus different kernels would respond to a rotated version of the same feature and the correspondence amongst feature maps' activations in  $N_l$  dimension gets lost.

Gram matrices  $G^l$  and  $A^l$  capture the spatial correlations of feature maps  $F^l$  and  $P^l$ , the presence of identical features inside  $I_s$  and  $I_c$  results in the correspondences of Gram matrices activations as well. Similarly, if both  $I_s$  and  $I_c$  do not possess identical features or the features exhibit rotational variation then the activation pattern in Gram matrices changes. In Figure9, we present a simple real world example of natural images which shows how the variations in global scenes affect the correspondence between the activations of  $F^l$  and  $P^l$  and thus the activation pattern of Gram matrices  $G^l$  and  $A^l$ . We visualize the Gram matrices of the feature maps from the layer *conv3\_4* of VGG-19. We specifically chose this intermediate layer as it is sensitive enough to intermediate features (neither too abstract nor too low level). We can see that the activation pattern of the Gram matrices corresponding to the style image and the content image 1 look very similar, even though the content image shows the variation in translation. Whereas, the activation pattern of the Gram matrix corresponding to content image 2 looks completely different even though the image has the same content as that of the style image. Such phenomenon is inevitable since the convolutional kernels are translational invariant and not rotationally invariant. Thus a small change in orientation of the pen instigates different kernels.

Since each entry of the Gram matrix is a summed up value of spatially correlated feature maps, therefore for a more detailed analysis we look into the feature maps corresponding to the peak activations of style image's Gram matrix, shown in red and green boxes. We also show the corresponding feature maps of both content images to which they get mapped to during optimization. The feature maps of the style image and the content image 1 show huge correlations in their activations. On the other

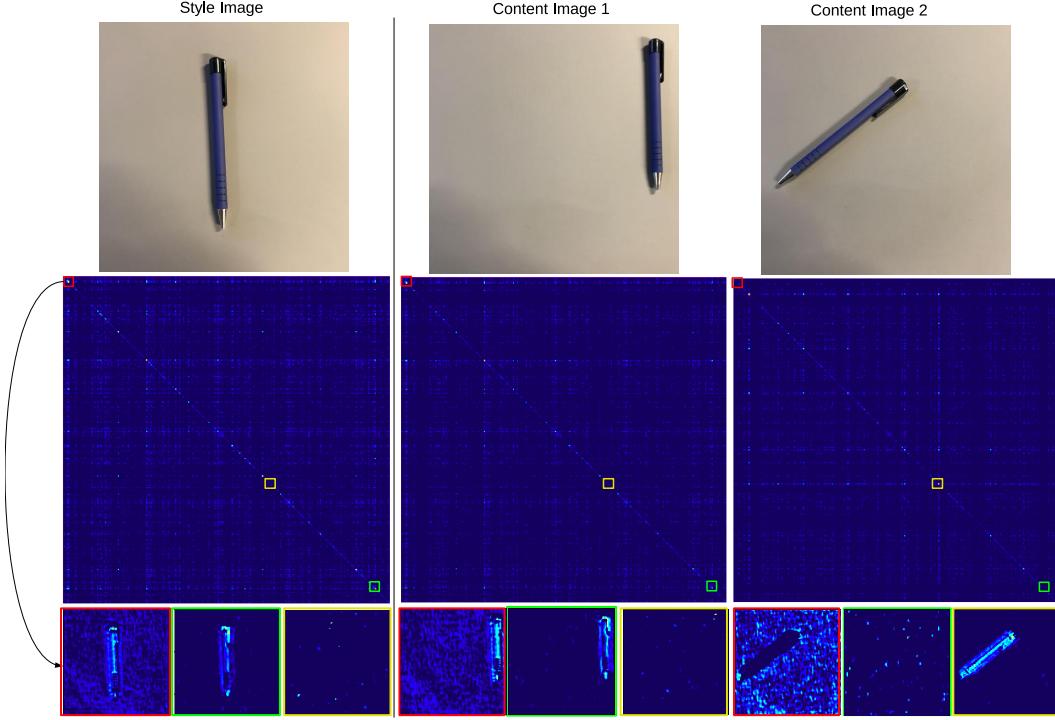


Figure 9: The top row shows the style and content images, whereas their corresponding Gram matrices are shown in the middle. The bottom row shows the feature activation maps corresponding to the peak activations in the Gram matrices. The feature maps in red and green boxes correspond to the top 2 activations in style image Gram matrix and the respective feature maps of both content images at those locations. The yellow box corresponds to the peak activation in content image 2’s Gram matrix. View electronically for better analysis.

hand, we see very weak correlations in the corresponding feature map activations of the style image and the content image 2. The yellow box corresponds to the peak activation in content image 2’s Gram matrix. The corresponding feature map (in yellow box) of the style image, to which it gets mapped to, contains very less information.

### 9.1.1 Mapping of Features in Gram Matrices

While optimizing for  $\mathcal{L}_{texture}$  in equation 1, we reduced the mean squared error between these Gram matrices values. Each entry of the Gram matrix represents the summed up statistics of feature maps’ correlations, where we have already lost the spatial information. Thus the optimization only enforces the summed up statistics on each entry of Gram matrix and not the spatial constraints. In Figure9, using the Gram matrix structure, different feature maps correlations of  $I_s$  and  $I_c$  get mapped to each other. We observe that this mapping plays a huge role in the hallucination of textures in the content images.

Figure10 shows how the feature representation of the reference style image in Figure9 are enforced on the content image’s feature maps during optimization. For the content image 1, there is huge mismatch in the activation levels of the style image feature maps and the corresponding feature maps of the content image 1. The top row shows how the information is enforced, while the bottom row shows how the information is diminished in the content image 1’s features. As explained earlier, the Gram matrix entries are just the summed up feature correlations values, having no spatial information thus the texture get hallucinated randomly across the image. The VGG-19 kernels themselves remain unchanged during optimization, therefore an image search is performed in order to match the feature representations of the style image. This mismatch in the mappings of Gram matrix values explain how a small variation like rotation can induce big changes in the texture transfer. Similarly if we look at the evolution of feature maps of content image 2, then we see high correlation with the activation levels of the style image.

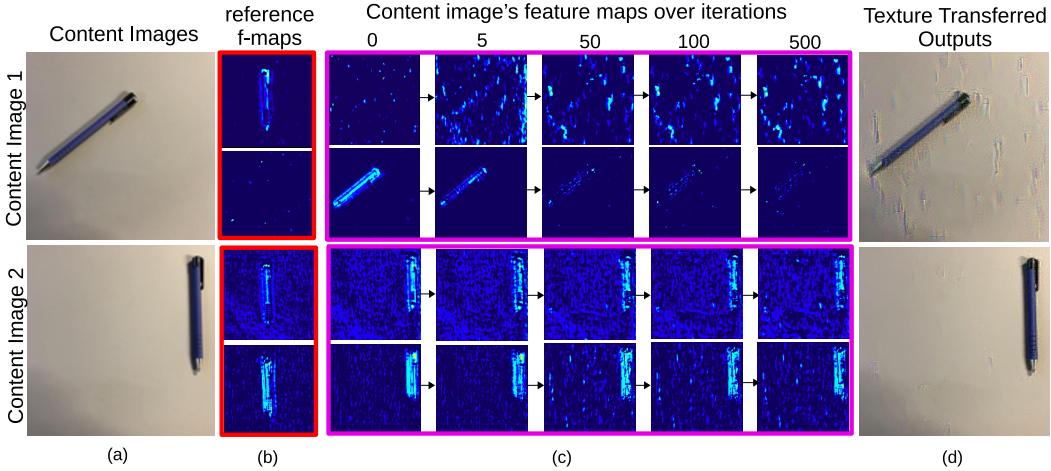


Figure 10: The left image (a) shows the initial content images which are first  $4\times$  bi-cubically downsampled and then upsampled back (b) shows the reference feature maps of style image from Figure9, the top image refers to the location of peak activation in style image's Gram matrix and the bottom one refers to the content image's. The middle figure (c) shows the evolution of respective feature maps of the content images. (d) shows the resultant image.

In style transfer [1], the content and the style image have different features which result in the huge mismatch of their Gram matrices activation pattern, and consequently the feature maps mappings. This mismatch results in the hallucination of new textures.

### 9.1.2 Exploiting Translational Invariant Mapping of Textures

We saw that the convolutional kernels extend their translational invariant property to the Gram matrices based texture transfer. In the following we discuss the efficiency and the implications of such texture mapping. In Figure 11(a), the content image is first bi-cubically down-sampled 4 times

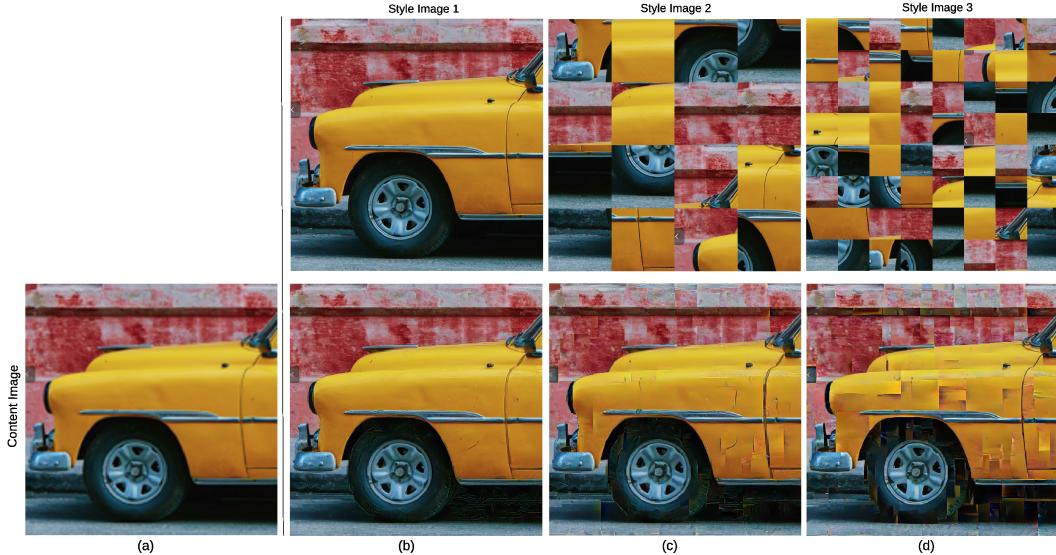


Figure 11: The left image (a) shows the content image which is  $4\times$  bi-cubically downsampled and then up-sampled again bi-cubically. (b) shows the texture transfer from style image 1. (c) and (d) shows the texture transfer from style image 2 and 3. Notice that the style image 2 and 3 have randomly shuffled patches of  $160 \times 160$  and  $80 \times 80$  respectively. Even then the Gram matrices have been able to satisfactorily map the textures correctly. Zoom in for texture analysis.

and then up-sampled back to the original dimensions bi-cubically. This is done to make sure that there is enough difference in the textures of the image pair which needs to be recovered via Gram matrices based optimization.

Figure 11(b) demonstrates the texture transfer from the style image 1. The existence of the identical features in both images enabled  $\mathcal{L}_t$  to map the texture from the style image correctly. Figure 11(c) shows the texture transfer from a relatively complex style image which has randomly shuffled patches of size 80 x 80. Both content and style images have similar features (ignoring the new edge pattern in the style image). Now according to the generally known behavior of Gram matrices based texture transfer the texture from the style image should not follow any spatial constraint and get transferred to the content image in haphazard fashion. However, we observe that the VGG-19 convolutional kernels are unreasonably effective in capturing the identical features correctly and then Gram matrix composition provides a very convenient structure in their translational invariant mapping. Thus the transfer of texture respects the global spatial arrangement of the content image. We do observe some artifacts, mainly because of the new edge structure in the style image and also because of the existence of homogeneous regions in the content image where finding the relevant features for mapping is difficult.

## 9.2 Model Agnostic Behavior.

Both TSRN-G and TSRN-S used relatively deep architectures for learning the mappings between  $I_{LR}$  and  $I_{HR}$ . To test the effectiveness of texture based super resolution, we also trained with a relatively shallow model having 6 residual blocks and the number of convolutional kernels reduced from 64 to 32. This reduced model is named TSRN-shallow, where the network's parameters are drastically reduced from  $\approx 1.07M$  to  $\approx 195k$  and the required memory storage from 4.3MB to 832.7 kB. Surprisingly, we did not observe much qualitative differences in the results, shown in Figure 12. However, the LPIPS metric evaluated TSRN-S and TSRN-G to be superior than TSRN-shallow, shown in Table 6. TSRN-shallow is trained the same way, that is pre-training with MSE for 10 epochs and then training with texture loss for another 100 epochs.

Metric	Set 5		Set 14		BSD 100		Urban	
	AlexNet	SNet	AlexNet	SNet	AlexNet	SNet	AlexNet	SNet
Bicubic	0.1585	0.1202	0.1731	0.1320	0.1463	0.1007	0.1552	0.1238
TSRN-S	0.0273	0.0394	0.0438	0.0483	<b>0.0478</b>	0.0420	0.0397	0.0404
TSRN-G	0.0285	<b>0.0358</b>	0.0463	<b>0.0456</b>	0.0481	<b>0.0404</b>	<b>0.0385</b>	<b>0.0392</b>
TSRN-shallow	0.0287	0.0397	<b>0.0388</b>	0.0473	0.0483	0.0406	0.0431	0.0415

Table 6: Comparison for 4x SISR on pre-trained AlexNet-linear and SqueezeNet-linear LPIPS metric[2]. Lower score is better.

## 9.3 Semantically Controlled Fast Style Transfer.

In this section, we demonstrate the effectiveness of our semantically controlled texture transfer scheme in training a convolutional neural network architecture. We use the same architecture and loss formulation as provided by Johnson et al. [3]. To train the networks for spatially controlled texture transfer, we used ADE20k dataset [46]. We made semantic division of the images into three classes i.e. sky, grass and others. Using the semantic masks provided with the dataset, we combined the classes of trees, fields, grass and plants into the “grass” class. The sky class remained the same as provided with the dataset whereas the rest of the classes were grouped in “others” class. Using the same methodology of semantic style transfer as explained in the main paper, we enforce different style images for these segmented regions. Figure 13 depicts the effectiveness of our spatially controlled fast style transfer scheme for multiple style images.

## 9.4 Architecture.

The architecture for texture based super resolution network (TSRN) is inspired by [9]. Similar form of architectures are used by many entries of NTIRE 2017 competition[26] and also by [4]. The

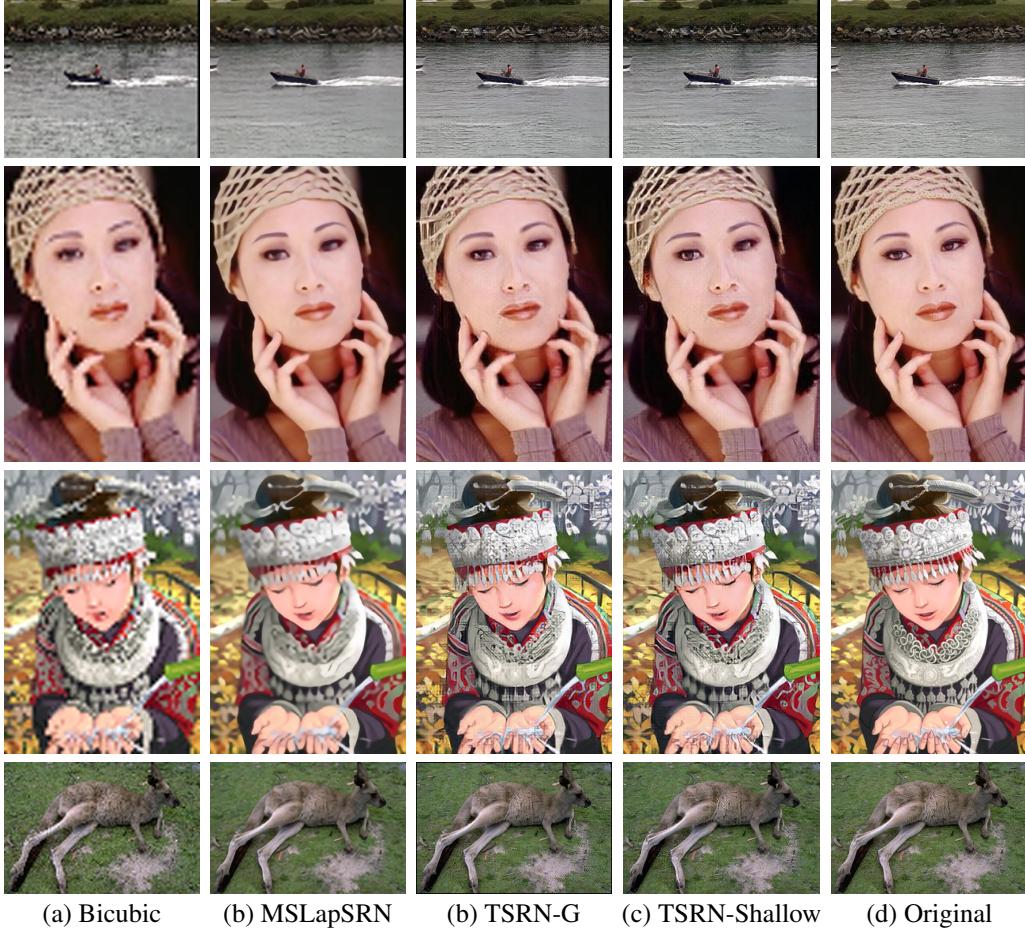


Figure 12: Visual Comparison to analyze the model agnostic behavior of texture based  $4 \times$  SISR. See that TSRN-Global and TSRN-small has very small visual differences. Zoom in for better analysis.

pictorial depiction of the architecture is shown in Figure 14. It must be noted that the architectures for the implementation of both global (TSRN-G) and semantically (TSRN-S) transferred texture methods has 10 residual blocks in each. This depth in the architecture is found to be enough to capture the texture mapping between a low resolution image  $I_{LR}$  and its high resolution counterpart  $I_{HR}$ .

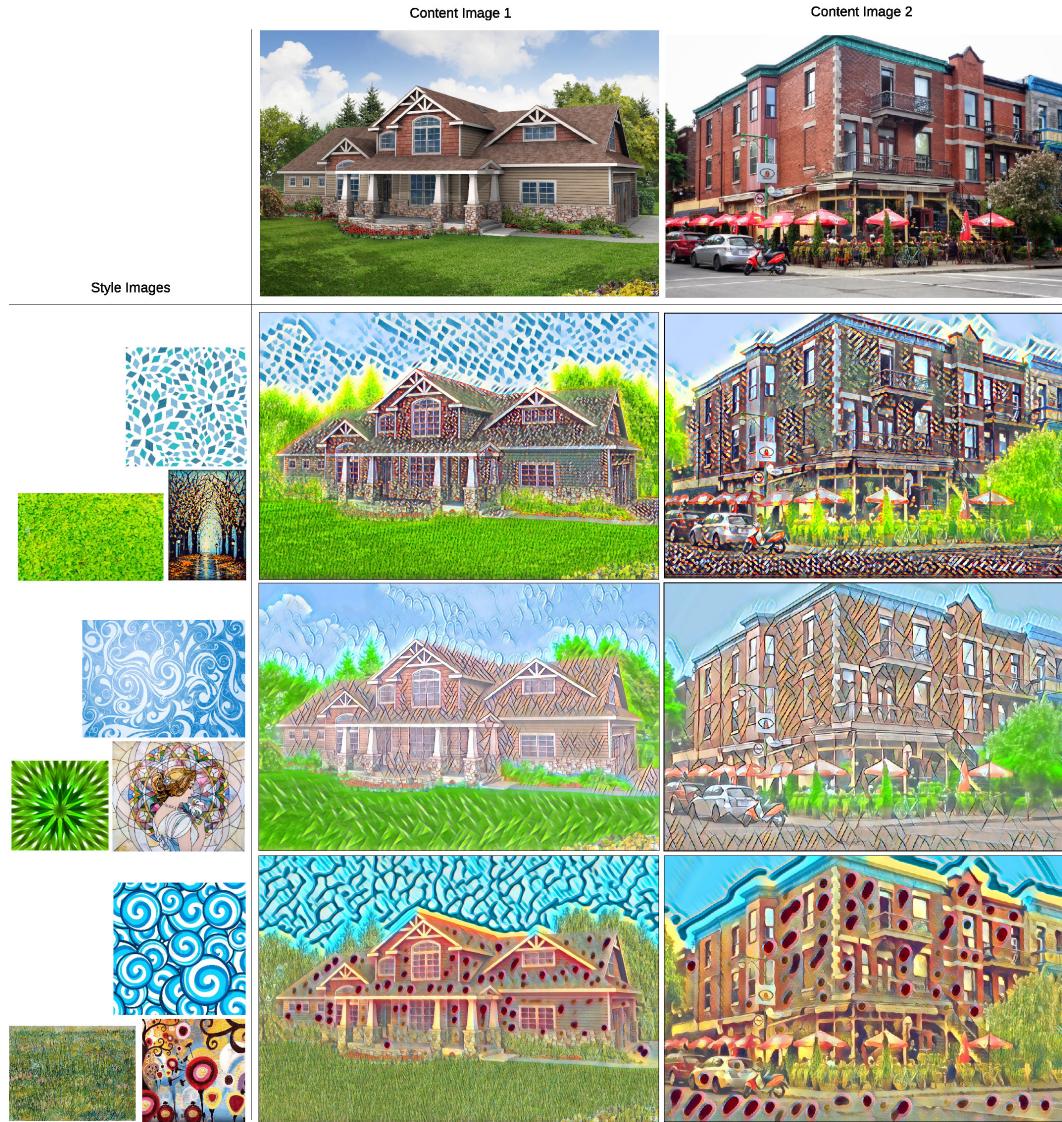


Figure 13: To show the effectiveness of our semantically controlled fast style transfer scheme. The images on the left show multiple style images which are semantically constrained to the sky, green and “others” regions of the content images 1 and 2 respectively.

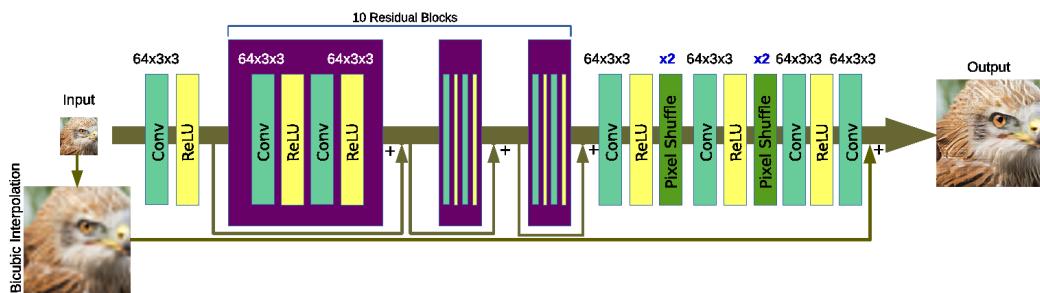


Figure 14: Pictorial illustration of the TSRN architecture used in the paper.