Enhancing Perceptual Loss with Adversarial Feature Matching for Super-Resolution

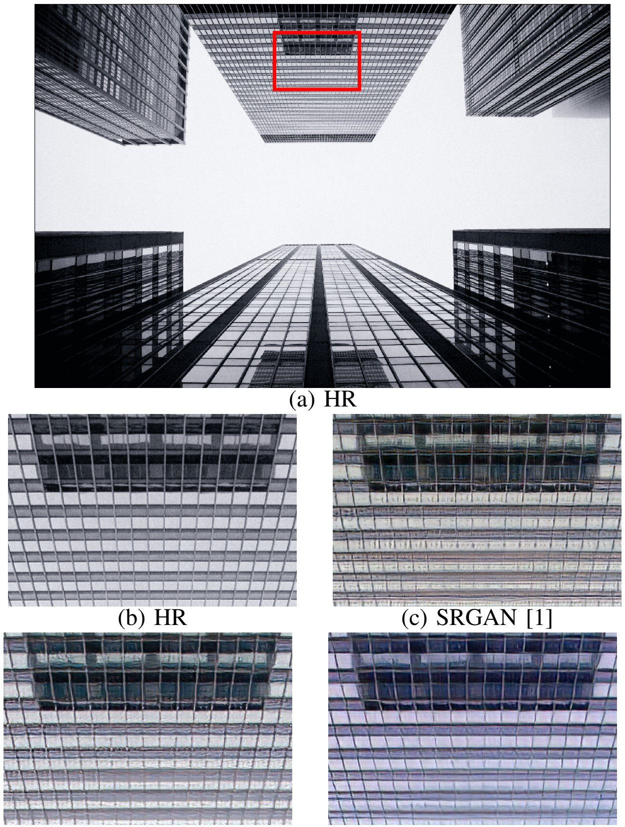
Akella Ravi Tej[†], Shirsendu Sukanta Halder[*‡], Arunav Pratap Shandeelya[*§], Vinod Pankajakshan[†]
[‡] *Carnegie Mellon University, USA*
[†] *Indian Institute of Technology Roorkee, India*
[§] *International Institute of Information Technology Bhubaneswar, India*
Email: ravitej.akella@gmail.com

***Abstract*—Single image super-resolution (SISR) is an ill-posed problem with an indeterminate number of valid solutions. Solving this problem with neural networks would require access to extensive experience, either presented as a large training set over natural images or a condensed representation from another pre-trained network. Perceptual loss functions, which belong to the latter category, have achieved breakthrough success in SISR and several other computer vision tasks. While perceptual loss plays a central role in the generation of photo-realistic images, it also produces undesired pattern artifacts in the super-resolved outputs. In this paper, we show that the root cause of these pattern artifacts can be traced back to a mismatch between the pre-training objective of perceptual loss and the super-resolution objective. To address this issue, we propose to augment the existing perceptual loss formulation with a novel content loss function that uses the latent features of a discriminator network to filter the unwanted artifacts across several levels of adversarial similarity. Further, our modification has a stabilizing effect on non-convex optimization in adversarial training. The proposed approach offers notable gains in perceptual quality based on an extensive human evaluation study and a competent reconstruction fidelity when tested on objective evaluation metrics.**

***Index Terms*—Single Image Super-Resolution, Perceptual Loss Functions, Generative Adversarial Networks**

I. Introduction

High-resolution (HR) images are perceived as more visually-pleasing than their corresponding mappings in low-resolution (LR) space since they form a better illusion of continuity. This perceived greater utility of HR images over LR images places a growing demand for signal processing techniques that learn a mapping between the HR and LR spaces. More formally, the problem of generating an HR image from several LR images is referred to as super-resolution reconstruction.



(a) HR

(b) HR

(c) SRGAN [1]

(d) EnhanceNet [2]

(e) *Ours* ($M_{pc\sigma va}$)

Fig. 1: Demonstration of pattern artifacts introduced by perceptual loss. Comparing our method with state-of-the-art SR models that use perceptual loss on DIV2K (super-resolved images are zoomed in for better comparison).

Our approach deals with a sub-problem of SR, where an HR image needs to be reconstructed from a single LR image, commonly known as *Single Image Super-Resolution* (SISR). Since SISR is an ill-posed inverse problem with multiple valid HR outputs for a single LR input, modern supervised learning approaches [3]–[5] restrict their solution space by learning a strong prior. For their capacity to learn complex and abstract representations, deep convolutional neural networks (CNNs) possess a favorable inductive bias for learning this prior. While CNNs trained on point-estimate loss functions attain state-of-the-art performance on peak signal-to-noise ratio (PSNR) metric, the generated images are overly smooth and have an unnatural appearance [1]. Further, this problem worsens at high upscaling factors, causing a steep drop in the generation quality of SR images. To enforce photo-realism in the generated images, recent methods use perceptual loss [2], [6] and adversarial loss [1] as objective functions for modeling the high-dimensional multi-modal distribution of natural HR images. While generative adversarial networks (GANs) have shown great potential in the generation of visually-realistic

[*] Equal contribution

images, their non-convex loss landscape results in unstable training. This severely limits current approaches that apply adversarial loss in combination with point-estimate loss or perceptual loss to ensure stable training. On the other hand, pre-trained perceptual loss functions provide a stable restoration of lost high-frequency components, although they often also introduce undesired artifacts in the generated outputs that current approaches fail to eliminate [6].

In this paper, we highlight that the pre-training objective of perceptual loss does not match with the true super-resolution objective. As a consequence, perceptual loss additionally transfers the biases from its pre-training stage that surfaces as pattern artifacts in the generated images. To address this objective mismatch, we propose a novel content loss [1] formulation that is an ensemble of content losses derived from the convolutional layers of a discriminator network. Each layer of the discriminator learns a unique abstraction for differentiating between the real and generated images, thereby allowing our content loss to address the removal of pattern artifacts that are identifiable across numerous levels of adversarial similarity. Further, we show that the proposed approach has connections with previously proposed techniques for stabilizing adversarial training in GANs. Finally, we conduct an extensive mean-opinion-score (MOS) test and the standard objective evaluation to demonstrate the gains in perceptual quality and content-preservation in the generated images.

II. Background and Related Work

A. Single Image Super-Resolution (SISR)

SISR involves the reconstruction of an *HR* image while limiting the contextual information to a single *LR* image. Early SISR approaches relied on filtering and interpolation [7], generating overly smooth images. Example-based approaches address this issue by learning a strong prior from internal similarities in the same image [8] or by externally learning a mapping between the *LR* and *HR* patches. With sufficient data, external example-based approaches can be effectively implemented in standard supervised learning frameworks like sparse-representation coding [9] and dictionary-based learning [10].

The recent success of deep CNN architectures propelled Dong *et al.* [3] into using a 3-layer CNN for SISR, which subsequently gave rise to a new direction of SR research with improved training methodologies. Kim *et al.* [4], [5] showed that recursive convolutions and residual learning could be used to realize deeper architectures that perform significantly better than shallow networks. To control the parameter count as we explore deeper architectures, Tai *et al.* [11] formulated SR as a recursive learning task. The use of recursive residual blocks vastly reduces the model parameters, enabling fast and efficient training of substantially deeper CNN models. Lim *et al.* [12] combined the ideas of multi-scale reconstruction and residual learning to achieve superior *HR* reconstruction at high upscaling factors.

B. Perceptual Quality

Despite vast advances in the architecture design of CNNs, the use of point-estimate loss functions (e.g., mean squared error) consistently gave rise to blurry images [5]. This is because point-estimate loss functions suffer from regression-to-the-mean problem at high upscaling factors. In other words, an optimal point-wise estimator returns the mean of many valid *HR* interpolations, resulting in blurry *HR* images. Another line of work that has attracted a lot of attention is the design of objective functions that focus on high-level image semantics over pixel-level details. The path taken by these approaches broadly falls into two classes: (i) directly emphasizing high-level feature reconstruction by optimizing in the latent feature-space of a pre-trained network (i.e., perceptual loss [6]), (ii) iteratively pushing the distribution of generated SR images closer to the distribution over natural *HR* images using a discriminator network (i.e., adversarial training [1], [13]).

Johnson *et al.* [6] was the first to introduce a perceptual loss in SR, by using the high-level features of an ImageNet [14] trained VGG network [15] to obtain sharp, visually pleasing images. The problem of SR was also explored in the context of adversarial learning by Ledig *et al.* [1]. Further, this approach also uses perceptual loss for efficient reconstruction of finer *HR* details. Taking inspiration from Gatys *et al.* [16], Sajjadi *et al.* [2] proposed a texture-matching loss in addition to adversarial and perceptual losses for the reconstruction of high-level texture details in *SR* images. These approaches further show that the PSNR metric used to measure the reconstruction fidelity in SISR tasks correlates poorly with the human perception of image quality. In their experiments, a network trained using *MSE* achieves a high PSNR score but fails to generate visually pleasing outputs relative to a network trained on perceptual loss or adversarial loss. To address this issue, Wang *et al.* [17] proposed a 2-stage training framework, a PSNR-oriented training followed by GAN-based fine-tuning, for trading-off fidelity with perceptual quality.

III. Proposed Method

Perceptual loss functions have several properties that make them appealing in the context of SR, *viz.* (i) they do not suffer from regression-to-the-mean problem like point-estimate loss functions, (ii) the CNN-based architecture of the pre-trained network makes them more stable to local deformations in the *LR* image, and (iii) they demonstrate a lower variance for stationary textures in the input, which are abundant in natural images. In other words, perceptual loss functions are low variance estimators that can produce stable high-frequency content and, consequently, sharp output images. However, perceptual loss functions were originally trained for a classification task on the ImageNet dataset [14], which makes them sensitive to the differentiating texture patterns observed across the 1000 training classes. Thus, using a pre-trained perceptual loss to optimize an SR model often causes unwanted texture patterns in the generated images.

Our main aim is to efficiently filter out the artifacts introduced by the pre-trained perceptual loss function. We

[1] Some works in the literature refer to content loss as feature matching loss.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

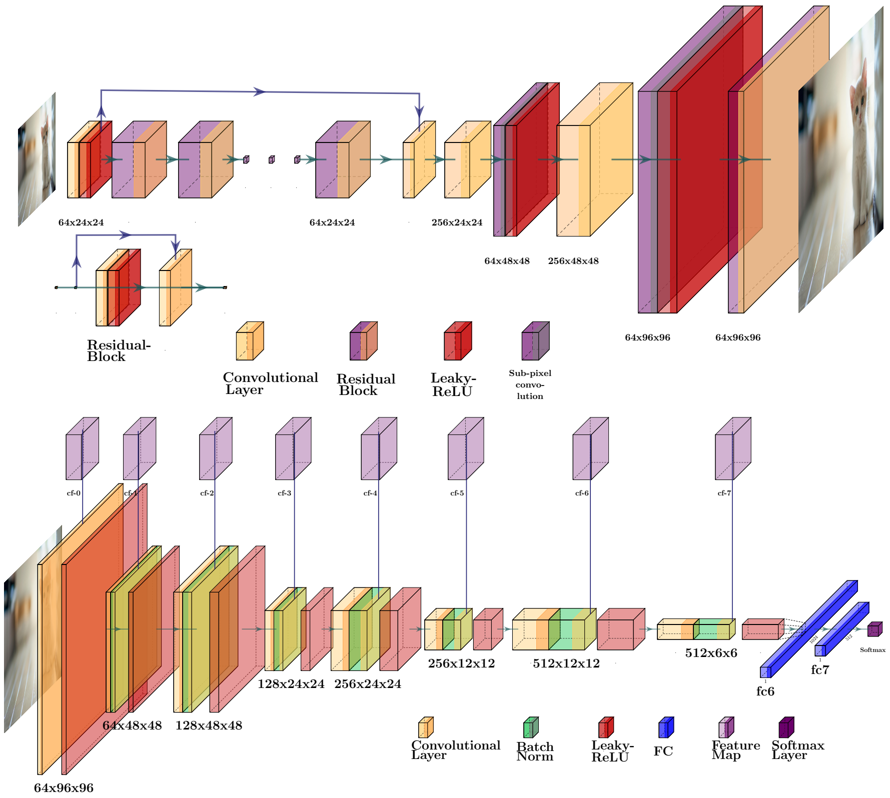


Fig. 2: *Top*: Architecture of our generator network. *Bottom:* Discriminator architecture with feature maps from the *Conv* layers (in purple) that are used in the computation of our content loss.

accomplish this by introducing a new content loss formulation that extends the traditional adversarial training framework. Unlike prior SR approaches that only use the final layer of a discriminator network, we derive our content loss from all its latent *Conv* layers. As a result, our approach provides stronger supervision for the generator's training while also stabilizing non-convex optimization in GANs (more details can be found in Sec. IV). Unlike perceptual loss, the proposed content loss matches the true SR objective for the following reasons:

- The training data for the discriminator, i.e., the proposed content loss network, is sampled from the distribution implicitly modeled by the generator network itself. In contrast, perceptual loss [6] network was originally trained for a discriminative task on the ImageNet dataset.
- The discriminator network adaptively learns features that are most discriminative of generated SR outputs versus natural *HR* images. Thus, the feature space of a discriminator network offers a natural choice of statistics for the generator to match. Crudely speaking, the proposed content loss advocates photo-realism, a sub-goal of SR.

A. Network Architecture

We use the SRGAN [1] architecture in all our experiments, to have a fair comparison with prior works.

Generator: A fully-convolutional feed-forward network comprising of an encoder and a decoder module. The encoder consists of a stack of $N = 16$ identical residual blocks. Each residual block consists of two *Conv* layers with 3×3 kernels, 64 channels and a LeakyReLU ($\alpha = 0.2$) activation function. The up-sampling decoder consists of two sub-pixel convolutional layers [18], each increasing the resolution by a factor of 2×. In contrast to the original SRGAN architecture, we avoid using Batch-Normalization layers in the generator because of its insensitivity to changes in input statistics, leading to unwanted artifacts and limited generalizability [17].

Discriminator: This network comprises of 8 *Conv* layers using a 3×3 kernel of stride length 1 and 4×4 kernel with stride length 2 in an alternating fashion. The number of channels in these *Conv* layers increases linearly from 64 to 512 as we go deeper into the architecture. There exists a Batch-Normalization layer and a LeakyReLU ($\alpha = 0.2$) activation between every two *Conv* layers. The last *Conv* layer is followed by 2 *Dense* layers and a *Sigmoid* neuron that outputs the final probability. The complete architecture is displayed in Fig. 2.

B. Objective Function

We formulate the overall SR objective ($\mathcal{L}$) as a weighted combination of the following loss functions:

$$\mathcal{L} = \mathcal{L}_{content} + \lambda \mathcal{L}_{adv} + \eta \mathcal{L}_{point} + \gamma \mathcal{L}_{vgg} \qquad (1)$$

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Point-estimate loss focuses on the reconstruction of low frequency components in the generated SR images. Unlike previous approaches that exclusively use $L1$ [12] or $L2$ loss [2], [4] for this purpose, we use Huber loss [19], a hybrid of $L1$ and $L2$ losses. Huber loss provides a robust loss function for regression that is less sensitive to outliers than $L2$ loss and more stable than $L1$ loss. Huber loss is defined as

$$\mathcal{L}_{point} = \begin{cases} \frac{1}{2}||I_{est} - I_{HR}||^2, & \text{if } |I_{est} - I_{HR}| < 1. \\ |I_{est} - I_{HR}| - 0.5, & \text{otherwise,} \end{cases} \quad (2)$$

where the low-resolution images, estimated super-resolution images and target high-resolution images are denoted by I_{LR}, I_{est} and I_{HR} respectively. Both the $L1$ and $L2$ losses in Eq. 2 also include averaging over all the image dimensions, which is not explicitly written for simplicity.

Perceptual loss [16] computes the squared $L2$ norm between the target HR and the output SR images in the latent feature space of a pre-trained VGG network [15]. Optimizing with perceptual loss instead of pixel-wise losses constrains the generator to produce images that match the high-level feature representations of the target images, thereby enforcing the reconstruction of high-frequency components in HR space. Let ψ_i denote the i^{th} feature layer of the $VGG19$ network. $\mathcal{L}_{vgg}$ is defined as:

$$\mathcal{L}_{vgg} = ||\psi_i(I_{est}) - \psi_i(I_{HR})||^2 \quad (3)$$

Adversarial loss [13] directly optimizes for photo-realism. The adversarial framework involves the joint-training of two networks, a generator G and a discriminator D. The generator loss is defined as the negative log-probability of discriminator for the generator's outputs.

$$\mathcal{L}_{adv} = \mathbb{E}_{I_{LR}}[-\log(D(G(I_{LR})))] \quad (4)$$

D is optimized over an opposing objective $\mathcal{L}_D$ to differentiate the generated images I_{est} from the target images I_{HR}.

$$\mathcal{L}_D = \mathbb{E}_{I_{LR}}[-\log(1 - D(G(I_{LR})))] + \mathbb{E}_{I_{HR}}[-\log D(I_{HR})] \quad (5)$$

Content loss extends the standard adversarial loss as the squared $L2$ norm across all the latent $Conv$ feature maps of the discriminator network D for I_{HR} and I_{est} images. Since the layers in D learn a hierarchy of differentiating representations of real and fake images, we optimize over an ensemble of content losses derived from all the $Conv$ feature maps. We do not consider dense features for this purpose as they lose spatial information, limiting their utility in the reconstruction of high-frequency components. Further, we use pre-activated features for computing the content loss as the activation layer sparsifies the feature maps and consequently weakens discriminator supervision [17].

Let ϕ_i denote a function that returns the pre-activated feature maps corresponding to the i^{th} $Conv$ block of D. Then, the i^{th} content loss is defined as

$$\mathcal{L}^i_{content} = ||\phi_i(I_{est}) - \phi_i(I_{HR})||^2 \quad (6)$$

Since there generally exist several $Conv$ layers in D with each layer learning a unique abstraction for differentiating between the natural HR and generated SR images, we wish to optimize over all the content losses simultaneously. While simply averaging over all the layer-wise content losses provides satisfactory results, such a scheme would not equitably optimize over all the content losses. This is because different content losses have varying optimization landscapes and a fixed weighted-averaging scheme would often provide an overall gradient that disproportionately favors only a subset of content losses. Moreover, selecting the weight of each loss over the course of training is also a non-trivial task.

To address this issue, we propose **softmax reweighing**, a dynamic mechanism to select the weight of each layer-wise content loss such that they are equitably optimized. Before starting the training, we re-weight the individual content losses to bring them to a comparable scale. During training, we rescale the gradient of each content loss by the softmax average of the content loss itself. Thus, during any update, the softmax averaging favors the optimization of a content loss with greater value over one with a smaller value. The overall content loss $\mathcal{L}_{content}$ is formulated as,

$$\begin{aligned} \nabla_\theta \mathcal{L}_{content} &= \sum_i \left\{ \frac{e^{\mathcal{L}^i_{content}}}{\sum_j e^{\mathcal{L}^j_{content}}} \right\} \nabla_\theta \mathcal{L}^i_{content} \\ \mathcal{L}_{content} &= \sum_i \left\{ \frac{e^{\mathcal{L}^i_{content}}}{\sum_j e^{\mathcal{L}^j_{content}}} \right\} \mathcal{L}^i_{content} \end{aligned} \quad (7)$$

where θ are the parameters of G and $\{.\}$ prevents gradient back-propagation (we do not compute the gradient for the softmax operation). In practice, we found this trick to evenly optimize over all the content losses and subsequently provide a notable improvement in the generator's performance.

IV. Connections with Prior Work

Since perceptual loss functions are derived from the latent features of a pre-trained network, a lot of its properties can be traced back to its pre-training strategies and the ImageNet dataset. More specifically, Geirhos *et al.* [20] showed that the features extracted by ImageNet-trained CNNs are sensitive to the texture patterns in the images. Complementary to our analysis, they investigate training strategies to obtain better feature extractors that are more robust to texture patterns and better match the human perception of image quality.

Using the intermediate layers of a discriminator for deriving a content loss function improves the stability of adversarial training [21]. This prevents the generator from over-training on the output statistics of a discriminator, while also encouraging it to model the multi-modal distribution of natural HR images. As a result, our content loss formulation provides the generator with greater supervision from the discriminator, and subsequently encourages the convergence of GANs, i.e., finding the Nash equilibrium of the minmax game.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

V. Experiments

A. Training Details

Data Preparation: Our training data consists of 800 high-quality images from the DIV2K [22] train set and 2650 high-quality images from the Flickr2K [23] dataset. All the SR experiments are conducted for a 4× scale factor between LR and HR images, i.e., 16× increase in image pixels. We extract random LR patches of spatial dimension 24×24, which correspond to HR patches of size 96×96. The initial training data is augmented with 90° rotations, horizontal and vertical flips. We observe that making the input data zero-centered (i.e., subtracting with the mean of the entire training dataset) provides an improvement in the generator's performance. For testing, we use 4 standard benchmark datasets: Set5 [24], Set14 [25], BSD100 [26] and Urban100 [27].

Training Parameters: We choose the MSRA initialization [28] for the weights of our network but further multiply them by 0.1, as reduced variance in the initial weights helps with faster convergence. Initial learning rate is set to 1e-4 for both the networks and reduced by a factor of 10 after every 200 epochs. Adam optimizer [29] is used to update the generator and discriminator networks. The weights (λ, η, γ) of $\mathcal{L}_{adv}, \mathcal{L}_{point}$ and $\mathcal{L}_{vgg}$ are set to 0.005, 0.01 and 0.5 respectively. All the SR models are trained for 2×10^5 updates with a batch size of 16.

B. Experimental Design

We train our SR network with a few variations in the overall loss function to systematically investigate the effect of different loss components on the generated outputs. We investigate the following combinations of loss components with their corresponding trained SR models:

- $\mathcal{L}_{point} \rightarrow M_p$
- $\mathcal{L}_{point} + \mathcal{L}_{vgg} + \mathcal{L}_{adv} \rightarrow M_{pva}$
- $\mathcal{L}_{point} + \mathcal{L}_{content} + \mathcal{L}_{adv} \rightarrow M_{pca}$
- $\mathcal{L}_{point} + \mathcal{L}_{content}(softmax) + \mathcal{L}_{adv} \rightarrow M_{pc\sigma a}$
- $\mathcal{L}_{point} + \mathcal{L}_{content}(softmax) + \mathcal{L}_{vgg} + \mathcal{L}_{adv} \rightarrow M_{pc\sigma va}$

Further, we also compare the reconstruction fidelity and perceptual quality of our final model with other SR models.

C. Quality Metrics

Mean Opinion Score (MOS): This metric assesses the perceptual quality of the generated images by gathering opinion scores from human raters. We use this metric to compare SR models trained with different combinations of loss components, thereby examining the influence each loss component on the perceptual quality. For the MOS test, we asked 25 human raters to assess the perceptual quality of the images with an integral score of 1 (low perceptual quality) to 5 (high perceptual quality). The human raters were first calibrated with 5 examples of Nearest Neighbor (NN) (score: 1) and HR images (score: 5). Subsequently, each human rater was given 8 versions of 20 randomly-presented images from BSD100: NN, bicubic, M_p, M_{pva}, M_{pca}, $M_{pc\sigma a}$, $M_{pc\sigma va}$, and HR (ground-truth). In other words, each of the 100 images (including its 8 versions) from the BSD100 dataset received a score from 5 human raters.

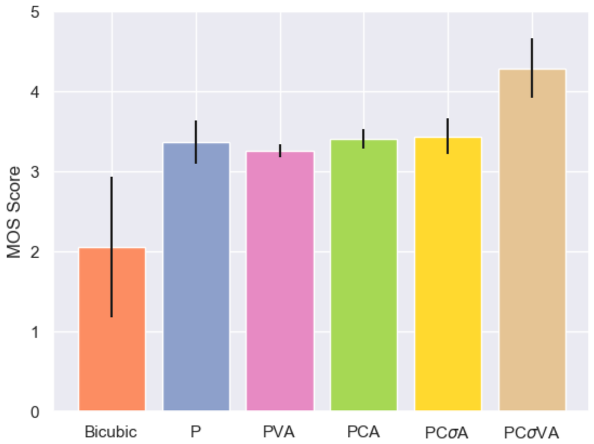


Fig. 3: The MOS scores for our model trained with different combinations of loss components.

Objective Evaluation: We report the peak signal-to-noise-ratio (PSNR), a standard evaluation metric in super-resolution for measuring the content preservation, i.e., reconstruction fidelity. Since PSNR correlates poorly with the human perception of image quality, we also report the performance on structural similarity (SSIM), and visual information fidelity (VIF) metrics. More specifically, SSIM uses spatial correlation, contrast distortion, and luminance masking to estimate the image quality. On the other hand, Visual Information Fidelity (VIF) uses natural scene statistics (NSS) in addition to an image degradation system and a human visual system (HVS) model for image assessment.

VI. Results and Analysis

A. Quantitative Analysis

The results from the MOS test are displayed in Fig. 3. We observed that the ratings for identical images did not show much variance and a majority of the users rated NN and HR images as 1 and 5 respectively. The results in Fig 3 indicate that $M_{pc\sigma va}$ model significantly outperforms our other models, with an average improvement of over 1 MOS score. Further, the MOS scores for M_{pva} model are inferior to $M_{pc\sigma va}$, confirming our hypothesis that the proposed content loss brings the best out of the perceptual and adversarial losses. Moreover, removing the perceptual ($M_{pc\sigma a}$) and adversarial loss components (M_p) results in similar performance degradation, which suggests the complementary nature of these loss components and highlights the importance of their presence for attaining a superior perceptual quality.

Table I summarizes the performance on objective evaluation metrics (PSNR, SSIM and VIF) for different combinations of loss components. M_p consistently outperforms all our other models since the point estimate loss directly optimizes for reconstruction fidelity. It also explains why M_p falls behind our other models in terms of MOS scores and $M_{pc\sigma va}$ (our model with highest perceptual quality) only delivers a modest

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE I: Quantitative Comparison of PSNR/SSIM/VIF values on test datasets for different combinations of loss components.

Dataset	*Bi-cubic*	M_p	M_{pva}	M_{pca}	$M_{pc\sigma a}$	$M_{pc\sigma va}$
Set5	28.42/0.810/0.443	**31.77/0.890/0.575**	23.49/0.848/0.505	**30.69/0.875/0.529**	**30.76/0.882/0.554**	30.03/0.864/0.520
Set14	26.00/0.704/0.380	**28.40/0.778/0.472**	21.88/0.725/0.403	**27.47/0.76/0.425**	**27.55/0.761/0.446**	26.74/0.742/0.418
BSD100 [26]	25.96/0.669/0.364	**27.46/0.732/0.422**	21.81/0.674/0.347	**26.86/0.714/0.390**	**26.72/0.715/0.402**	26.17/0.693/0.372
Urban100 [27]	23.15/0.659/0.371	**25.78/0.776/0.446**	20.82/0.715/0.371	**24.80/0.754/0.393**	**24.97/0.765/0.412**	24.40/0.744/0.380

TABLE II: Peak Signal-to-Noise-Ratio (PSNR) values of different SR methods on test datasets.

Dataset	Bi-cubic	DRCN	DRRN	DSRN	EDSR	E-Net	ESRGAN	LapSRN	SelfExSR	SRCNN	SRGAN	VDSR	*Ours*	*Ours*
		[5]	[11]	[30]	[12]	[2]	[17]	[31]	[27]	[3]	[1]	[4]	$M_{pc\sigma va}$	M_p
Set5	28.42	31.54	**31.68**	31.40	**32.64**	28.57	30.47	31.74	30.34	30.08	29.92	31.35	30.03	**31.77**
Set14	26.00	28.12	**28.21**	28.07	**28.95**	25.77	26.29	28.26	27.55	27.27	26.57	28.03	26.74	**28.40**
BSD100	25.96	27.23	**27.38**	27.25	**27.80**	24.93	25.32	27.43	26.85	26.7	25.5	27.29	26.17	**27.46**
Urban100	23.15	25.13	25.44	25.08	**26.86**	23.54	24.32	**25.51**	24.82	24.14	24.39	25.18	24.40	**25.78**

TABLE III: Structural Similarity (SSIM) values of different SR methods on test datasets.

Dataset	Bi-cubic	DRCN	DRRN	DSRN	EDSR	E-Net	ESRGAN	LapSRN	SelfExSR	SRCNN	SRGAN	VDSR	*Ours*	*Ours*
		[5]	[11]	[30]	[12]	[2]	[17]	[31]	[27]	[3]	[1]	[4]	$M_{pc\sigma va}$	M_p
Set5	0.810	0.885	**0.889**	0.883	**0.900**	0.81	0.852	**0.889**	0.863	0.853	0.851	0.882	0.864	**0.890**
Set14	0.704	0.769	0.772	0.770	**0.790**	0.678	0.698	**0.774**	0.755	0.743	0.709	0.770	0.742	**0.778**
BSD100	0.669	0.723	0.728	0.724	**0.744**	0.626	0.65	**0.731**	0.711	0.702	0.652	0.726	0.693	**0.732**
Urban100	0.659	0.751	0.764	0.747	**0.808**	0.693	0.733	**0.768**	0.739	0.705	0.731	0.753	0.744	**0.776**

TABLE IV: Visual Information Fidelity (VIF) values of different SR methods on test datasets.

Dataset	Bi-cubic	DRCN	EDSR	E-Net	ESRGAN	SelfExSR	SRCNN	SRGAN	*Ours*	*Ours*
		[5]	[12]	[2]	[17]	[27]	[3]	[1]	$M_{pc\sigma va}$	M_p
Set5	0.443	**0.540**	**0.574**	0.44	0.502	0.502	0.48	0.49	0.520	**0.575**
Set14	0.380	**0.418**	**0.452**	0.346	0.376	0.398	0.377	0.37	**0.418**	**0.472**
BSD100	0.364	0.361	**0.387**	0.293	0.313	0.346	0.333	0.303	**0.372**	**0.422**
Urban100	0.371	**0.380**	**0.452**	0.332	0.377	0.365	0.325	0.36	**0.380**	**0.446**

🟩 Best score
🟦 2^{nd} best score
🟪 3^{rd} best score

performance on objective evaluation metrics. In summary, none of the objective evaluation metrics from our experiments correlates with the human perception of image quality, i.e., MOS scores. Further, the superior reconstruction fidelity of $M_{pc\sigma a}$ to M_{pca} can be attributed to the softmax reweighing. Another interesting observation is that M_{pca} outperforms M_{pva}, suggesting that the proposed content loss (derived from the latent features of a discriminator network) can be a good proxy for the perceptual loss (derived from the latent features of a pre-trained VGG network).

The results from Table II, III, and IV suggest that the objective evaluation metrics favor models trained on point estimate loss functions, i.e., EDSR [12] and M_p (trained with Huber loss only). More interestingly, all the perceptually-motivated approaches (e.g., $M_{pc\sigma va}$, SRGAN [1], EnhanceNet [2]) are outperformed by simple models such as SRCNN [3] trained with point-estimate losses. This supports our previous conclusion that objective evaluation metrics correlate poorly with perceptual quality. Nevertheless, $M_{pc\sigma va}$ still provides a competent reconstruction fidelity relative to other SR methods.

B. *Qualitative Analysis*

From Fig. 4, it can be seen that using just the per-pixel loss as in M_p causes the output to blur. The use of perceptual and adversarial losses in M_{pva} results in a sharper image over M_p, although the image also tainted with square-like patterns, giving it with an artificial look. In general, adding perceptual loss increases the sharpness of super-resolved images, as visible from M_{pva} and $M_{pc\sigma va}$. Replacing perceptual loss with the proposed content loss in M_{pca} results in a smoother image with much fewer high-frequency artifacts. The use of softmax reweighing in $M_{pc\sigma a}$ provides a much cleaner image and removes any residual artifacts but still over-smoothens the final output. Finally, $M_{pc\sigma va}$ provides the most perceptually-convincing images with adequate frequency textures and intricate details. Interestingly, removing the perceptual loss causes a noticeable degradation in the perceptual quality for $M_{pc\sigma a}$, emphasizing on the importance of perceptual loss. This qualitative analysis is in coherence with the quantitative analysis of MOS scores.

Fig. 5 displays the outputs of SR methods trained on point-estimate loss functions along with $M_{pc\sigma va}$ in the increasing order of perceptual quality. In contrast to $M_{pc\sigma va}$, the other SR methods lack sharpness and fine details, which can be attributed to regression-to-the-mean problem of point-estimate losses. When compared to perceptual SR methods (see Fig. 1), $M_{pc\sigma va}$ produces cleaner images with fewer artifacts. More specifically, the building image of SRGAN [1] contains several box-like artifacts near the edges. The EnhanceNet [2] model trained using a VGG-based texture loss has incongruous texture patterns in its super-resolved outputs. In contrast, our method does not contain any block artifacts or conflicting texture patterns. In other words, although perceptual loss is crucial for obtaining sharp images, we demonstrate that sharpness alone does not directly correlate with the perceptual quality. With increased discriminator supervision, the proposed

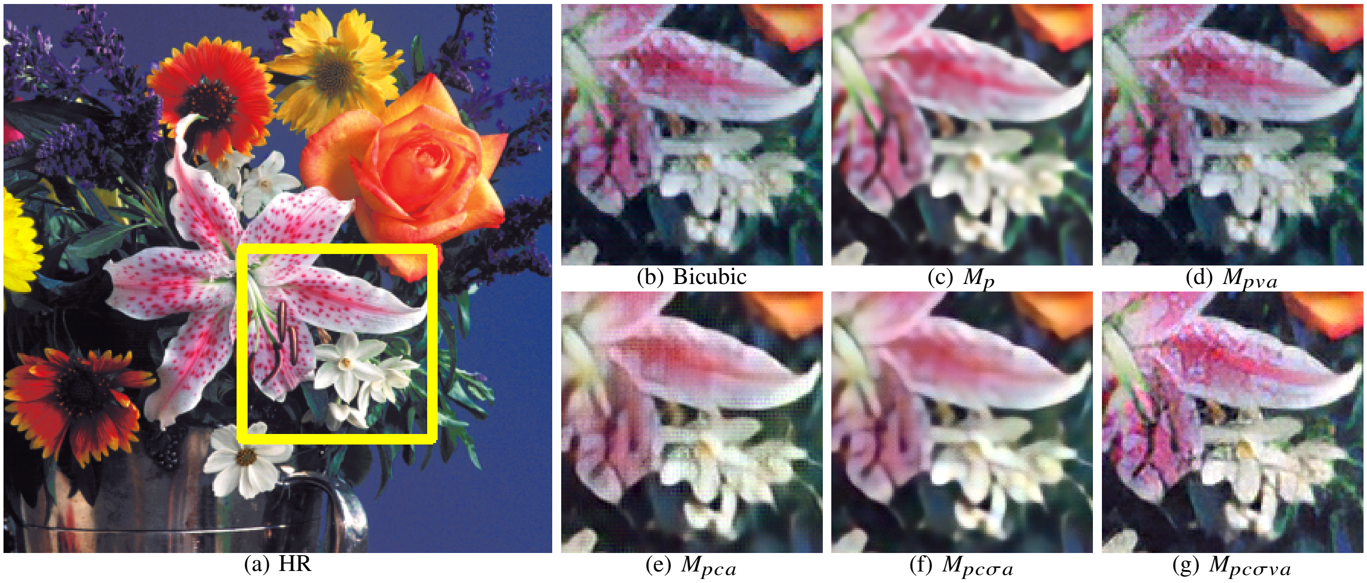

(a) HR (b) Bicubic (c) M_p (d) M_{pva} (e) M_{pca} (f) $M_{pc\sigma a}$ (g) $M_{pc\sigma va}$

Fig. 4: Qualitative comparison of outputs from different combinations of loss components on the *Flowers* image from Set14.

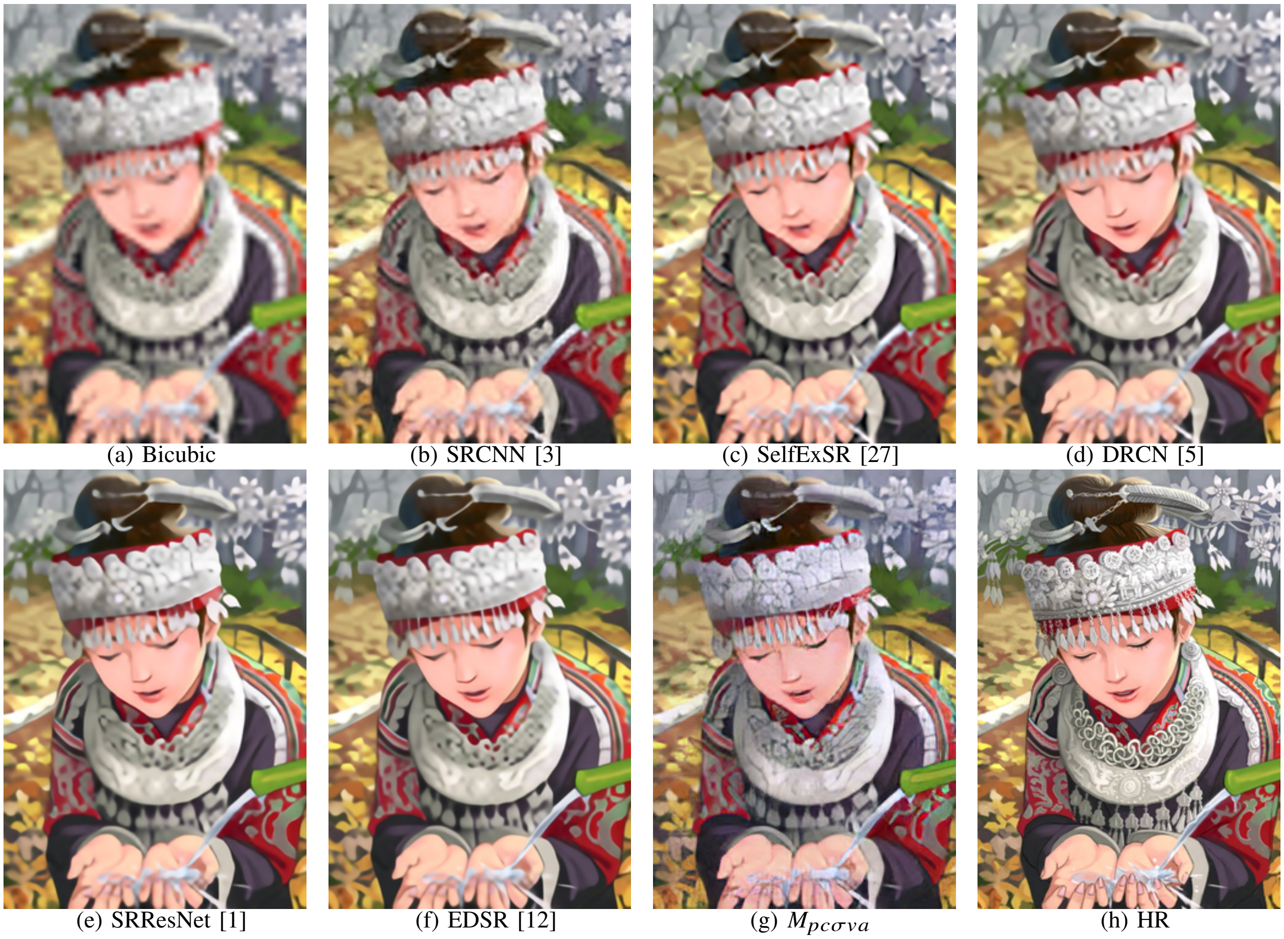

(a) Bicubic (b) SRCNN [3] (c) SelfExSR [27] (d) DRCN [5] (e) SRResNet [1] (f) EDSR [12] (g) $M_{pc\sigma va}$ (h) HR

Fig. 5: Qualitative comparison of our method ($M_{pc\sigma va}$) with other SR models on the *Comic* image from Set14.

content loss provides sharp SR images with pertinent high-frequency patterns.

VII. Conclusion

In this work, we investigated the challenges in training deep generative models with perceptual and adversarial losses for the super-resolution task. We showed that these loss functions have complementary advantages and can be effectively combined to overcome their individual shortcomings. We derived a novel content loss formulation from the latent features of discriminator network to (i) effectively eliminate the biases transferred from the perceptual loss, and (ii) stabilize adversarial training with increased supervision from the discriminator network. Further, we systematically studied the properties of the proposed content loss when combined with other loss functions. Our results confirm that the proposed content loss addresses the individual shortcomings of perceptual and adversarial losses, providing substantial gains in the perceptual quality of the generated images. Moreover, our approach also provides a competent reconstruction fidelity relative to state-of-the-art SR methods.

References

[1] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 105–114.

[2] M. S. M. Sajjadi, B. Schölkopf, and M. Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," in *Proceedings IEEE International Conference on Computer Vision (ICCV)*. Piscataway, NJ, USA: IEEE, Oct. 2017, pp. 4501–4510.

[3] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, Feb 2016.

[4] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[5] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 1637–1645.

[6] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.

[7] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of Applied Meteorology*, vol. 18, no. 8, pp. 1016–1022, 1979. [Online]. Available: https://doi.org/10.1175/1520-0450(1979)018¡1016:LFIOAT¿2.0.CO;2

[8] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *ICCV*, 2009. [Online]. Available: http://www.wisdom.weizmann.ac.il/ vision/SingleImageSR.html

[9] T. Peleg and M. Elad, "A statistical prediction model based on sparse representations for single image super-resolution," *IEEE transactions on image processing*, vol. 23, no. 6, pp. 2569–2582, 2014.

[10] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2216–2223.

[11] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 2790–2798.

[12] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[16] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[17] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[18] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 1874–1883.

[19] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 03 1964. [Online]. Available: https://doi.org/10.1214/aoms/1177703732

[20] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bygh9j09KX

[21] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. [Online]. Available: https://openreview.net/forum?id=Hk4_qw5xe

[22] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[23] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, B. Lim *et al.*, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[24] M. Bevilacqua, A. Roumy, C. Guillemot, and M. line Alberi Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2012, pp. 135.1–135.10.

[25] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[26] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, July 2001, pp. 416–423 vol.2.

[27] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5197–5206.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, "Image super-resolution via dual-state recurrent networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[31] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE