Understanding and Simplifying Perceptual Distances

Dan Amir and Yair Weiss
The Hebrew University of Jerusalem
{dan.amir,yair.weiss}@mail.huji.ac.il

Abstract

Perceptual metrics based on features of deep Convolutional Neural Networks (CNNs) have shown remarkable success when used as loss functions in a range of computer vision problems and significantly outperform classical losses such as L1 or L2 in pixel space. The source of this success remains somewhat mysterious, especially since a good loss does not require a particular CNN architecture nor a particular training method. In this paper we show that similar success can be achieved even with losses based on features of a deep CNN with random filters. We use the tool of infinite CNNs to derive an analytical form for perceptual similarity in such CNNs, and prove that the perceptual distance between two images is equivalent to the maximum mean discrepancy (MMD) distance between local distributions of small patches in the two images. We use this equivalence to propose a simple metric for comparing two images which directly computes the MMD between local distributions of patches in the two images. Our proposed metric is simple to understand, requires no deep networks, and gives comparable performance to perceptual metrics in a range of computer vision tasks.

1. Introduction

Following the enormous success of CNNs for object recognition tasks [37, 24] and evidence for the generality of their learned representations [11], Gatys et al. [15] suggested comparing images in the feature space of such networks' intermediate layers instead of in pixel space. They suggested that such representations are more sensitive to the semantic content of the image and less to the exact appearance of the objects in the image. Following their success, the use of perceptual losses has spread and showed promising results for a variety of tasks: ranging from image restoration tasks, such as super-resolution [20, 26], image deblurring [25] and image inpainting [28], image generation [12, 14] and image domain transfer [6]. Despite its universal success and applicability, perceptual losses suffer from several drawbacks:

- **Computational cost** - computing intermediate feature representations and propagating gradients backwards through a large object recognition CNN can be very expensive.
- **Domain specificity** - while representations learned on ImageNet [10] tend to transfer well to a range of computer vision tasks, such features may not be applicable for domains where the image statistics differ drastically from those in ImageNet.
- **Interpretability** - it is not well understood when, how and why perceptual losses succeed, and how to tune their hyperparameters. As evidence for the severity of this problem, one can observe the large inconsistency in the literature in the choice of most hyper-parameters for perceptual loss - the choice of specific layers (or combinations of them), whether features are extracted pre or post activation and whether to normalize activations prior to distance computation.

In this work we derive a better understanding of the reasons behind the success of the commonly used perceptual losses, and use this understanding to derive a simpler, well understood loss that can serve as an alternative. Specifically, we show that losses based on CNNs with *random* filters are almost equally "perceptual" and can serve as a suitable loss function in image prediction tasks. We then use the recent tool of infinite random networks to derive an analytical form for perceptual loss in such CNNs. Specifically, we prove that the perceptual distance between two images is equivalent to the maximum mean discrepancy (MMD) distance between local distributions of small patches in the two images. We use this equivalence to propose a simple metric for comparing two images which directly computes the MMD between the distribution of patches in the two images. Our proposed metric is simple to understand, requires no deep networks, and gives comparable performance to perceptual metrics in a range of computer vision tasks.

1.1. The Success of Perceptual Losses

The effectiveness of perceptual losses compared to standard losses is evidenced by two empirical successes: (1)

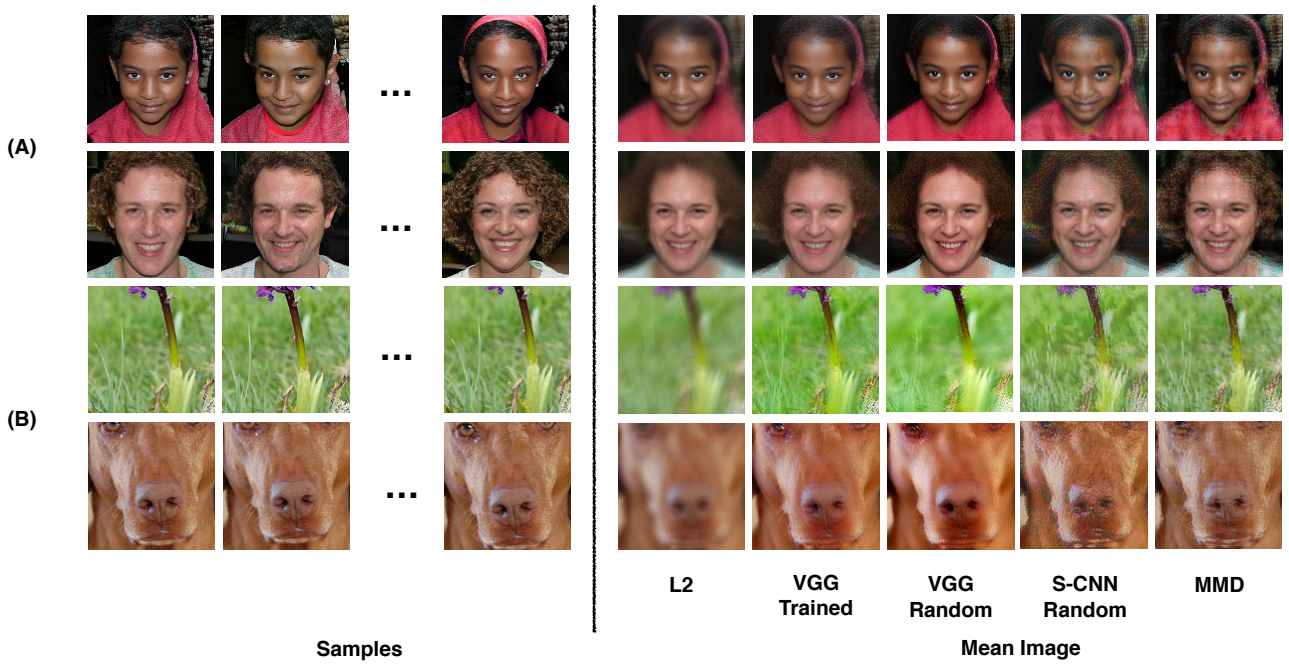

Figure 1. The generalized mean image problem. Given a set of target images (left), we find an image that minimizes the sum of losses with respect to all targets (right). In (A) sets were generated using StyleGAN as described in Section 5, and in (B) target images consist of slightly different crops of a single image from ImageNet. When the loss is L2, the optimal image is blurry , while both perceptual losses and our new, proposed loss, give sharp, realistic images.

when used as losses in tasks in which the targets are natural images they lead to sharp, realistic images while standard losses lead to blurry, non-realistic images (e.g. [3]) and (2) they better correlate with human judgements of patch similarity [40]. In this section, we illustrate these successes and also show the surprising utility of perceptual losses based on random filters in those two scenarios.

To investigate the effectiveness of perceptual losses in image prediction tasks, i.e. tasks in which the targets are natural images, we first consider why the mean squared error (MSE) over pixel values fails on such tasks. Consider for example, the task of image super-resolution in which for every low resolution (LR) image one needs to predict the correct high resolution (HR) image. A trained model which generalizes well will minimize the reconstruction loss with respect to the true distribution of HR images given LR images. Since such a model is deterministic, for a single LR image it should predict a single HR image even though many plausible HR images could have been mapped to the same LR image. The optimal model will choose the image that minimizes the average expected reconstruction loss over the distribution of possible HR images. When MSE is used, this is simply the mean. Since the mean of many sharp images with non-aligned textures and edges is a blurry image, the predictions of the model will be blurry. Thus, a sensible requirement from a loss function would be that the image which minimizes the expected loss over all probable target images is sharp and realistic.

To empirically investigate this property in controlled setting, we define the "Generalized Image Mean" (GIM) optimization problem, where given a set of target images $Y = \{y_1, ..., y_N\}$ and an image reconstruction loss $\mathcal{L}$ we seek to find an image $\hat{y}$ that minimizes the sum of the losses with respect to all target images:

$$\hat{y} = \arg\min_{y} \frac{1}{N} \sum_{i} \mathcal{L}(y, y_i) \tag{1}$$

For the MSE loss this reduces to the mean of the set Y and for any differentiable loss we can approximate it by optimizing the value of $\hat{y}$ directly, using Stochastic Gradient Descent (SGD). Figure 1 shows results for the GIM problem for different sets and losses. In the first two rows, Y is a set of samples generated by StyleGAN [22] for a small neighborhood of latent codes and in the last two rows these are slightly different crops of the same image. When the L2 loss is used, the optimal image is clearly blurry and non-realistic. However, when the loss function is the perceptual loss with a trained VGG network, the optimal image is

sharper and much more realistic. These illustrative results confirm previous reports (e.g. [3, 5]) and demonstrate that the effectiveness of perceptual losses can be observed even in simple prediction problems.

1.2. Unreasonable Effectiveness of Random Filters

In the work of Zhang et al. [40] randomly initialized networks were found to be as "perceptual" as low level losses and far less consistent with human judgements compared to trained CNNs. The judgements were based on a two alternative forced choice (2AFC) task, where each trial is composed of a random reference image patch and two random deformations of the patch. Subjects are asked to choose which of the two deformations is more similar perceptually to the reference patch. In contrast to their results, we found that by adding a few simple modifications, random networks can achieve comparable performance to learned networks on this task. First, to reduce the variance of predicted judgements, we compute each 2AFC judgement by initializing 20 random networks and using voting between all results. Second, we constrain the random filters in the first layer to have zero-mean, thus ignoring the DC component for every channel (by subtracting the mean of the randomly drawn filter). We evaluate random VGG16 networks using this scheme and obtain accuracy closer to that of supervised methods as can be seen in figure 2.

Motivated by the surprising success of these random networks, we turn to analyze a simpler CNN architecture that consists of the standard "ingredients" of common CNNs: convolutions and pooling layers, which we call the Simple-CNN (S-CNN). This architecture consists of a single spatial convolution with kernel size P followed by D 1×1 convolution layers and a single average pooling layer with window size W and strides S. All intermediate layers have constant width C and all convolutional layers are followed by ReLU activation. This architecture shares some similarities with previous works which investigated the use of limited local receptive fields and 1×1 convolutions [4, 34]. As shown in figure 2, random networks of this architecture with parameters $P = 3$, $D = 6$, $W = 32$, $S = 16$ and $C = 1024$ work significantly better than low-level metrics and slightly better than random VGG in the 2AFC experiment.

Figure 1 shows that random networks also work surprisingly well when used as loss functions for the GIM task. Again, we compute $\hat{y}$ using equation 1 where the loss is a "perceptual" loss with random CNNs. The computed image $\hat{y}$ is sharp and realistic, comparable to the predictions when a trained VGG is used for the loss. This is true both for a random VGG network and a random S-CNN.

In summary, our results reaffirm previous reports regarding the remarkable effectiveness of perceptual losses compared to standard losses. However, the success of random CNNs challenges the conventional wisdom that this success has to do with the fact that CNNs trained on image discrimination learn representations that are semantically meaningful. How, then, can we understand the success of perceptual losses?

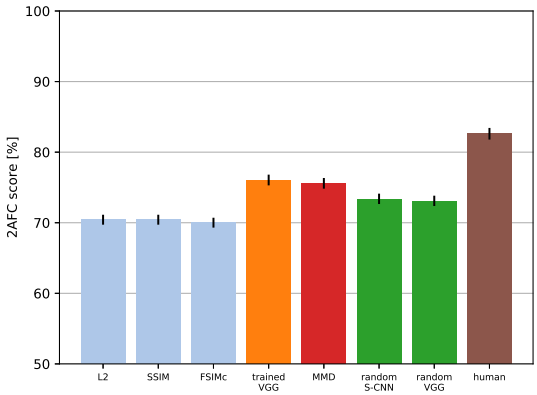

Figure 2. Test accuracy on the CNN-based and traditional image distortion 2AFC task from [40]. Trained CNNs, random CNNs and our suggested MMD loss, significantly outperform classic low-level distance metrics.

2. Analysis: Perceptual Similarity in Random S-CNNs Converges to MMD Between Patch Distributions.

Our main analytical result shows that perceptual similarity between two images using a random S-CNN converges to a distance between the local distribution of small patches in the two images. The distance between distributions is measured using Maximum Mean Discrepancy (MMD) [17]. Roughly speaking, given two distributions $P(x), Q(x)$, MMD measures the maximal difference between the expectations of a smooth function of x in the two distributions. We now give the formal definition of MMD and describe its properties before stating our result.

Definition: given two distributions P, Q, and a norm over functions $\|f\|_H$ whose reproducing kernel is K, the MMD between P, Q is given by:

$$MMD(P,Q) = \max_{\|f\|_H<1} E_P[f(x)] - E_Q[f(x)] \qquad (2)$$

Similar to the Wasserstein distance [1], the MMD between the distributions P and Q is given by a critic function $f(x)$. If two distributions are identical, then for any critic function we will have $E_P[f] = E_Q[f]$ and the more dissimilar the distributions, the larger the difference. The critic is constrained to have unit Hilbert norm in function space. For the commonly used RBF Kernel, $\|f\|_H$ penalizes the magnitude of the Fourier transform of f at different frequencies and thus measures the smoothness of f.

As shown by [17], under reasonable assumptions on the kernel K, the MMD distance between two distributions is zero if and only if $P = Q$. However, unlike the Wasserstein distance and other distances between distributions, it can be efficiently computed in high-dimensions.

Given N IID samples s_i from P and N IID samples t_i from Q, an empirical (biased) estimate of the MMD between P and Q is given by:

$$\hat{MMD}_K(S,T) = \frac{1}{N^2}\sum_{i,j} K(s_i,s_j) - 2K(s_i,t_j) + K(t_i,t_j)$$

Theorem 2.1. *L2 distance over normalized representations in the post-pooling layer of an infinite random S-CNN with patch size P and pooling window size W is equal to the average $\hat{MMD}_K$ distance between the distributions of patches of size P within windows of size W in the two images. The kernel $K(p_1, p_2)$ of two patches is a robust, monotonically decreasing function of the L2 distance between the two patches.*

Proof. The proof follows from a line of recent results regarding infinitely-wide CNNs [33, 8]. Given two images x_1, x_2, a neural network architecture and a layer h, these results allow us to calculate $K(x_1, x_2)$, defined as the dot product between $h(x_1)$ and $h(x_2)$ in an infinite CNN with that architecture. When computing perceptual similarity, we are measuring the L2 distance between representations which can be written as $\|h(x_1) - h(x_2)\|^2 = K(x_1, x_1) - 2K(x_1, x_2) + K(x_2, x_2)$. By using the form of K derived in previous work, we arrive at the result. For completeness, a full proof is given in the supplementary material. □

Corollary: If the set of possible outputs $\{y_i\}$ for a given input x are a sequence of images that all have the same distribution over patches of size P in all pooling windows of size W, then $\hat{y}$ defined by equation 1 using a perceptual loss with an infinite random S-CNN should have the same local distribution over patches of size P as each of the original images.

Proof: This follows directly from the fact that $MMD(P, Q)$ is zero if and only if $P = Q$.

While our theorem is for the simplified CNN with random weights and infinite width, similar results can be obtained for an S-CNN with learned weights and finite width: the perceptual distance is still the MMD between the two distributions, but the kernel between patches K is now a learned kernel which may not satisfy the conditions that ensure that the MMD is zero if and only if the two distributions are the same.

Our analysis suggests an alternative explanation for the success of perceptual losses in many settings. If all the training images have the same local distributions over patches, the predicted image will have the same distribution. Thus if all the training images have sharp gradients, we should expect the predicted image to have sharp gradients as well.

3. A New Loss

If indeed the success of perceptual loss is largely due to its minimization of the distance between distributions of local patches, we should be able to achieve similar success with a more direct loss. To test this hypothesis, we define a new loss that directly measures the average $\hat{MMD}_K$ distance between the distributions of patches of size P within windows of size W in the two images. Rather than using deep networks to define the kernel between two patches, we simply replace it with the standard Gaussian RBF kernel, defined as:

$$k(p_1, p_2) = \exp\left(-\frac{\|p_1 - p_2\|^2}{2\sigma^2}\right) \quad (3)$$

As mentioned above, the MMD distance between two distributions over patches $P(p)$ and $Q(p)$ with this kernel measures how different $E_P(f)$ can be from $E_Q(f)$, where f is constrained to be a smooth function of the patches.

Our MMD loss has four hyper-parameters: the bandwidth of the Gaussian kernel (σ), the size (W) and strides (S) of the pooling window and the patch size (P). In addition, one can decide whether or not to use channel normalization (in which the DC of the patch in each channel is ignored when comparing the two patches). All the results in this paper used patch size of 3, $W = 32$ and $S = 16$ corresponding to the post-pooling features of the S-CNN. We varied σ for different applications but it was always in the range $(0.5, 0.75)$. We used channel normalization only in the 2AFC experiments.

Computing the distance for all pairs of patches within every pooling window can become computationally intensive for large pooling windows. Therefore, we approximate the Gaussian kernel using Random Fourier Features [36]. Using this approximation, the MMD loss can be represented in any auto differentiation framework as a two layer CNN, where the first layer is a convolution with random weights $w \sim \mathcal{N}\left(0, \frac{1}{\sigma^2}\right)$ and biases $b \sim \mathcal{U}\left(0, 2\pi\right)$, followed by cosine activations and an average pool layer with a pooling window W and pooling strides S.

Our new loss function was motivated by the success of random S-CNNs in the GIM experiments and the 2AFC task of [40]. As can be seen in figures 1 and 2, our loss which directly computes distances between distributions of patches in the two images performs very well on both tasks. In the 2AFC task it is within the confidence interval of the pretrained VGG.

4. Related Work

Comparing two images by measuring the distance between their local histograms has a long history in Computer

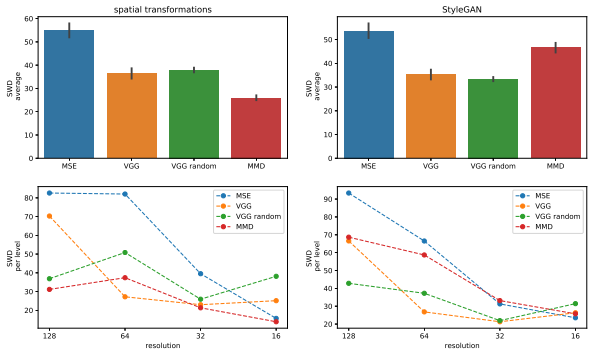

Figure 3. How well do different distance metrics preserve the statistics of target distribution? We compute the Sliced Wasserstein Distance over patches in different Laplacian Pyramid scales [21] and show that for both of our image optimization tasks, MMD and VGG (either trained or randomly initialized) outperform the MSE.

Vision (e.g. [16, 9]). Our analytical result shows that the modern way of comparing images (using perceptual distance in a CNN) is closely related to this classical idea. A major difference between the classical idea and the MMD approach is that the MMD distance does not require binning the patches into a discrete histogram and can work directly with the non-binned samples. Among the different works that have considered maintaining image statistics as a goal of image prediction, the closest to our work is that of [31, 30]. In their work, Mechrez et al. introduced a loss function called "Contextual Loss" that approximates the KL-divergence between the global distribution of local features in the two images. While this loss function can be applied directly to the pixel values of image patches, in practice, they focused on the use of such loss above a CNN based image representation. Therefore, it is framed mainly as a complementary method to the standard perceptual loss and not as a replacement as in our approach.

In recent years, many works have investigated the behavior of infinitely wide random CNNs [33, 27, 19] but mainly as a tool for analyzing the inductive bias of discriminative neural networks and their generalization capabilities. More closely related to our work, is that of Cheng et al. [7] which analyzed the prior induced by CNNs that output an image for a fixed noise input and utilized this prior for image restoration inference problems. To the best of our knowledge, our work is the first to utilize this analytic tool for better understanding of the use of CNNs as feature extractors for images comparison.

Previous works [38, 13] showed that for texture synthesis and style transfer, using the second order statistics of random networks is as successful as pretrained CNNs. He et al. [18] devised a data-dependent initialization scheme for VGG and showed comparable results to the pre-trained network on style transfer and texture synthesis tasks. In this context, we further support their findings and extend them to the general context of image prediction tasks. Bruna et al [5] compared feature representations of a pretrained VGG network and the hand-crafted scattering network (which represents average pooling of local features) as part of an optimization based method for image super-resolution where both representations performed comparably.

5. Experiments

In the first set of experiments, we re-examined the GIM problem (as in figure 1) but in a quantitative fashion. Specifically we repeatedly created datasets $Y = \{y_i\}$ of different possible outputs for a given input x and used equation 1 to solve for $\hat{y}$. We refer to each set Y as a problem instance and each problem instance was created using one of two scenarios:

Random Spatial Transformations - For every problem instance in this setting we first draw a random example from ImageNet [10]. This image is then resized such that its short side is of 256 pixels and the aspect ratio is preserved. Then, for each example, a random square crop of size 128 is drawn within a $[-4, 4]$ range from the center crop at each direction. Such uncertainty over y aims to approximate the type of uncertainty over high-resolution images observed in the super-resolution task or image deblurring with motion blur.

StyleGAN Synthetic Faces - We use StyleGAN [22], a state of the art GAN architecture, trained for generating samples of face images. For each problem instance, we first draw a single z code from the original predefined distribution and then draw N modulated latent codes from a small Gaussian around z and generate the set of N images that those codes are mapped to by the generator.

For every problem instance of the described problem settings, we sample 50 target examples. For every loss evaluated, we initialize the image with the mean of the 50 examples (equivalent to pretraining with L2 loss). The image is optimized with PyTorch [35] through SGD with the commonly used Adam optimizer [23] with default parameters. See the supplementary material for implementation details.

For quantitative evaluation we focus on the perceptual loss with pretrained VGG (referred to henceforth as VGG loss) and our suggested MMD loss. We generate 100 random instances for each of the two scenarios and compute the optimal predicted image for each loss once. Similar to the evaluation of image generation models, we mainly wish to examine how perceptive the results are, *i.e.* how well does the distribution of image means fits the distribution of input examples, and not how well the specific images are reconstructed. We use the Sliced Wasserstein Distance (SWD) over image patches in a Laplacian pyramid as introduced by Karras et al. [21] for GAN evaluation. We use the orig-

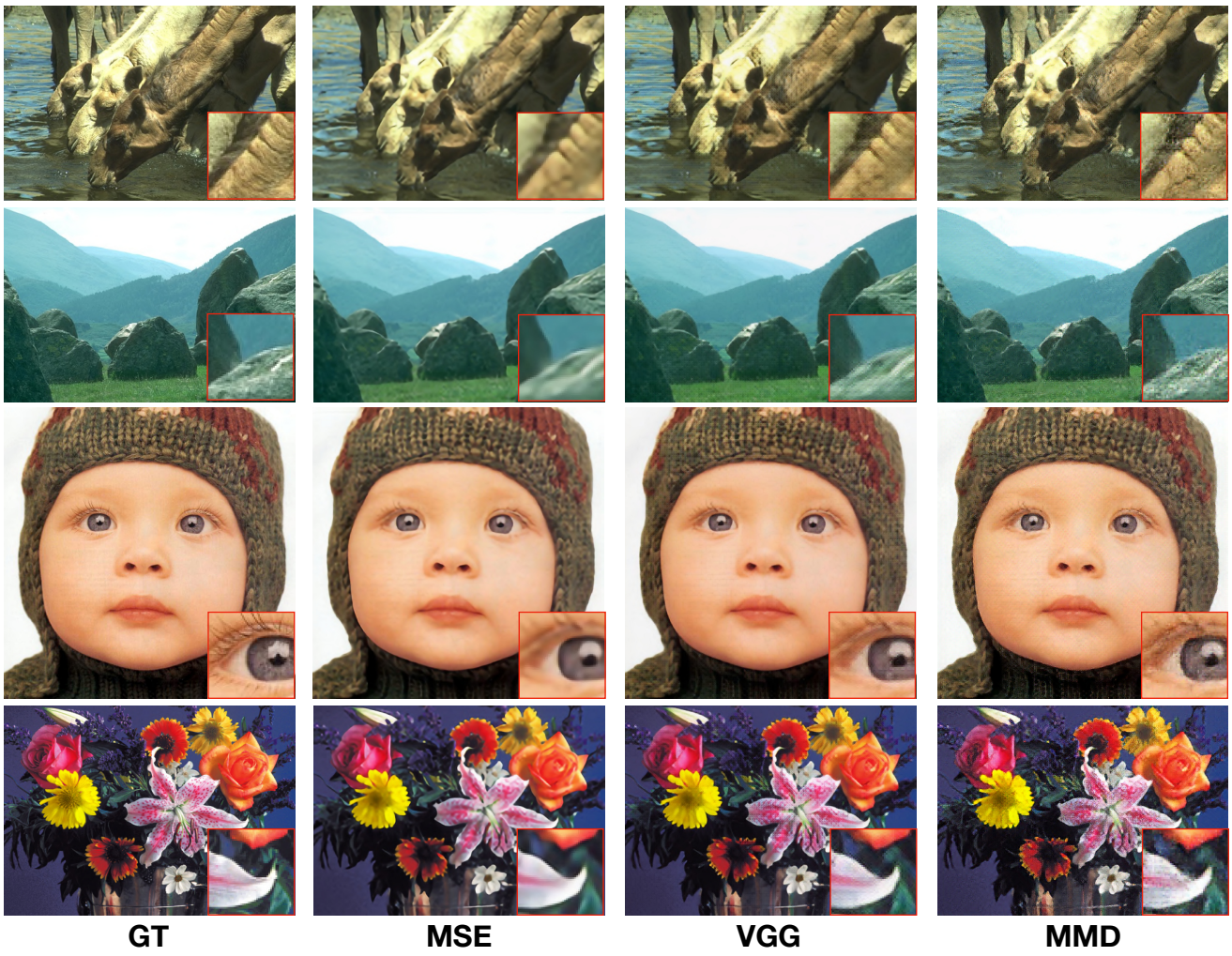

Figure 4. Qualitative comparison of image super-resolution results for different loss functions. Our new loss clearly produces sharper results than both MSE and VGG based perceptual loss, while also producing less artifacts than the VGG loss.

inal implementation [1] with default parameters. It should be noted that although both the MMD loss and SWD compare images using patch statistics, SWD uses larger patches and multiple scales. In general we found that SWD scores are consistent with our qualitative impression regarding the results (see additional results in supplementary material).

Figure 3 shows the results. Consistent with our hypothesis, *in both scenarios random VGG networks performed as well as trained VGG and our simple MMD was comparable to VGG.* All three losses performed significantly better than L2 and the largest gap as expected is observed in the higher pyramid levels (consistent with the qualitative observation that L2 reconstructions are blurry).

5.1. Super resolution

To investigate the performance of the MMD loss on real world problems we turn to the task of $\times 4$ image super resolution. Our implementation is based on the super-resolution demo from the official PyTorch examples package[2] with some modifications. See the supplementary material for implementation details.

We evaluate our models over the standard super resolution benchmark datasets: BSD100 [29], Set14 [39] and Set5 [2], and compare our results with those of the models from [26] obtained by different combinations of L2 distance on pixels, perceptual losses and adversarial loss. We recompute the evaluation metrics for these baseline methods based on their predicted images [3] . We compare the

[1] https : / / github . com / tkarras / progressive _ growing_of_gans

[2] https://github.com/pytorch/examples

[3] https://tinyurl.com/ammv9a37

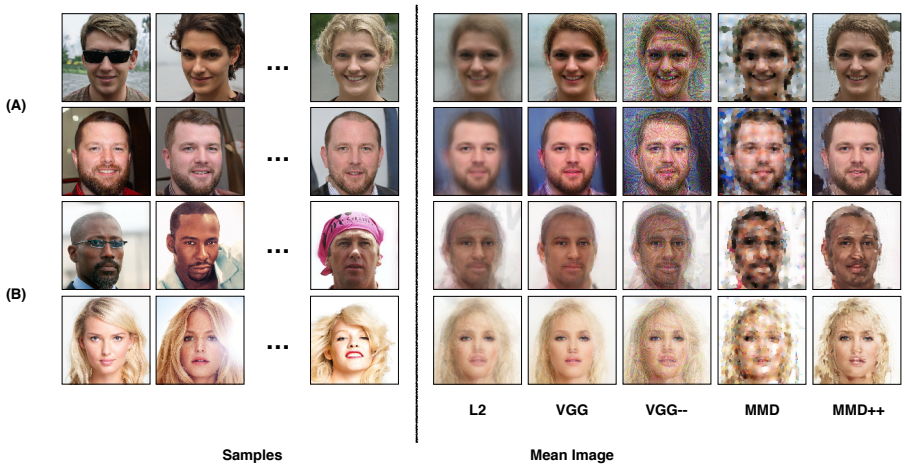

Figure 5. An example for a GIM tasks in which VGG succeeds and MMD fails. In (A) sets were generated using StyleGAN as described in Section 5 with larger variance, and in (B) target images consist of random CelebHQ image and its 9 nearest neighbors. While the MMD loss fails, adding a local term equivalent to pre-pooling layers of the S-CNN in MMD++ results in significant improvement. Replacing the multi-layer VGG loss with only the deepest of the five layers in VGG– also introduces very strong artifacts.

	Dataset	BSD100			Set5			Set14		
	Metric	NIQE	PSNR	SWD	NIQE	PSNR	SWD	NIQE	PSNR	SWD
Method	Loss									
Ground Truth		3.13	-	-	4.85	-	-	4.78	-	-
Bicubic		5.76	24.25	56.21	6.32	25.89	48.00	5.98	23.73	53.24
Our Method	MSE	5.38	26.05	47.71	6.40	29.61	39.17	5.28	**26.26**	56.84
	VGG	7.55	24.88	27.87	7.73	28.22	**23.81**	7.40	25.09	28.14
	MMD	**4.55**	23.54	**26.89**	**4.67**	26.38	24.40	**4.38**	23.61	**18.92**
SRGAN	MSE	5.27	**26.27**	48.23	6.08	**30.05**	43.42	6.86	25.57	55.94
	VGG_{2_2}	7.19	25.06	32.41	6.53	28.69	30.88	5.19	24.61	41.08
	Adversarial + MSE	3.78	24.45	30.47	4.10	28.68	31.76	**3.52**	24.30	41.68
	Adversarial + VGG_{2_2}	3.45	24.20	27.82	4.40	27.67	22.71	3.82	23.74	34.60
	Adversarial + VGG_{5_4}	**3.37**	23.70	**14.09**	**3.91**	27.23	**14.45**	3.66	23.43	**14.04**

Table 1. Summary of quantitative results on the super-resolution task for our method and relevant baselines on standard benchmark test sets. Methods involving adversarial losses appear in gray. Metrics include the no-reference image quality assessment method NIQE (lower is better), the standard PSNR and SWD. For each metric, both the best performing method and the best among non-adversarial methods appear in bold.

models using both reference and no-reference metrics. For the reference based evaluation, we use the standard PSNR. As a no-reference metric we use NIQE [32] which compares the statistics of large image patches to a prior model learned on natural scenes and was previously used to evaluate super resolution methods in [3] and SWD over three levels of the Laplacian Pyramid. We compute all metrics on the RGB images without any post-processing and use the standard MATLAB implementation of NIQE.

Predictions for some random examples appear in figure 4. It is clear that amongst our different models, the MMD model produces the sharpest results and while it produces some noisy high-frequency artifacts, they are minor compared to the VGG model. Quantitatively, it is clear that our MMD model, although relying only on low-level patch statistics, produces the highest quality results (in terms of NIQE score and SWD) between all non-adversarial methods. In contrast, our suggested loss function consistently

obtains the lowest PSNR across datasets, indicating that it encourages the network to distort the images further from the ground truth images (at least in terms of low level metrics). Comparing our and SRGAN's performance with MSE and VGG losses, it is clear that there is a slight positive trend towards SRGAN's implementation. This can be explained by the advantage of SRGAN's training scheme in terms of dataset size ($\times 35$), training iterations ($\times 8$) and architecture complexity (as explained in the supplementary material). In light of this, the MMD loss can potentially achieve better performance as more computationally intensive training should produce even better results - closer to the ballpark of adversarial training. These results are also in line with the results obtained in our GIM experiment for spatial transformations, where MMD produced the best results. Interestingly, both in our experiments and in the SRGAN work, training with only a combination of the VGG loss and MSE results in degradation in NIQE scores for all datasets compared to using only MSE loss. This may be due to the sensitivity of the NIQE score to high-frequency artifacts, as it mainly relies on the statistics of local image derivatives. While the VGG results, in both our configuration and SRGAN's implementation (with only a single VGG layer) show very strong periodic artifacts, it is possible that a different choice of layers would produce results closer in quality to those of the MMD loss. This again demonstrates the difficulty of working with an uninterpretable loss function - the space of different CNN-based perceptual loss configurations is large and once training fails, it is hard to determine what change can resolve it. Moreover, while in the MMD the parameters factorize the spatial aspects (patch size and pooling size) and the robustness to outliers (σ), our theoretical results suggest that for VGG, these are entangled, as the deeper the layer both the robustness and the spatial aspects increase.

6. Limitations and Extensions

While our results show that the performance of the MMD loss is comparable to that of the VGG loss in various settings, there are other settings where the behavior is quite different. This is particularly true when the losses are used to compare images that are highly dissimilar. figure 5 shows two examples for the GIM problem. In the top figures, we create sets of images from StyleGAN that have greater variability compared to figure 1 and in the bottom figures we create a set of 10 images Y by taking a random CelebHQ [21] image and its 9 nearest neighbors in the dataset. In both cases, finding the generalized mean using MMD gives highly nonrealistic images while VGG produces sharp images (see figure 5).

One way of understanding this failure of MMD, is that as shown in theorem 2.1, it is equivalent to a "perceptual loss" that is based on a *single layer* of an infinite random S-CNN. In contrast, losses based on VGG almost always use *multiple* layers. In fact, as shown in figure 5 when we use only the last layer of VGG (referred to as VGG- in the figure) we also obtain highly nonrealistic images. While optimizing the MMD loss by itself can be beneficial when the patch distribution of possible target images is similar, this is not necessarily the case when the patch distributions differ. In that case, an optimal image would match the distribution of patches randomly drawn from all target images (see analysis in supplementary material) which can explain the strong artifacts in figure 5.

It follows directly from the proof of Theorem 2.1 that for all layers other than the post-pooling layer, the equivalent perceptual loss is a robust loss over corresponding patches in the two images: $L(x, y) = 2 - 2\sum_i K(p_i, q_i)$ where q_i, p_i are ith patches in x, y respectively. Thus the top-left $P \times P$ patch in one image is compared to the top-left patch in the second image, and this is repeated for all patches at all locations. As shown in figure 5, a combination of a robust loss over patches with the MMD loss (referred to as MMD++ in the figure) produces much better images compared to pure MMD. Such a loss also reduces many of the periodic artifacts that are observed with the pure MMD loss in our experiments (see supplementary material). The supplementary material also shows that when the two images to be compared are more dissimilar, more differences are observed between losses based on trained VGG networks and losses based on random VGG networks, although most of these differences disappear when using trained weights only for the convolutions in the very first layer of VGG and random weights for all other layers.

7. Conclusion

As noted by previous authors, perceptual losses are remarkably effective in image prediction problems and lead to sharper, more realistic images when compared to standard losses such as MSE. Furthermore, they are much better than standard losses in predicting human similarity judgements. One explanation for this success is that CNNs trained on ImageNet learn semantically meaningful features. In this work, we have shown that CNNs with random filters can perform comparable to trained features on both tasks, which argues against the importance of semantically meaningful features as an explanation. As an alternative explanation, we used the tools of infinite random networks to show that perceptual losses in a CNN with random filters converge to the MMD distance between distributions of local patches in the two images. We then used this insight to devise a new loss which requires no pretrained deep networks and is simple to understand yet retains many of the successful properties of perceptual losses.

Acknowledgments: We thank the ISF, Israeli MOS, and the Gatsby foundation for financial support.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018.

[4] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

[5] Joan Bruna, Pablo Sprechmann, and Yann LeCun. Super-resolution with deep convolutional sufficient statistics. *arXiv preprint arXiv:1511.05666*, 2015.

[6] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017.

[7] Zezhou Cheng, Matheus Gadelha, Subhransu Maji, and Daniel Sheldon. A bayesian perspective on the deep image prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5443–5451, 2019.

[8] Youngmin Cho and Lawrence K Saul. Kernel methods for deep learning. In *Advances in neural information processing systems*, pages 342–350, 2009.

[9] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. volume 1, pages 886–893, 07 2005.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[11] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[12] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *arXiv preprint arXiv:1602.02644*, 2016.

[13] Len Du. How much deep learning does neural style transfer really need? an ablation study. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3150–3159, 2020.

[14] Aviv Gabbay and Yedid Hoshen. Demystifying inter-class disentanglement. In *International Conference on Learning Representations*, 2019.

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[16] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Efficient learning with sets of features. *J. Mach. Learn. Res.*, 8:725–760, May 2007.

[17] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[18] Kun He, Yan Wang, and John Hopcroft. A powerful generative model using random weights for the deep image representation. *arXiv preprint arXiv:1606.04801*, 2016.

[19] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

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

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[25] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018.

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

[27] Zhiyuan Li, Ruosong Wang, Dingli Yu, Simon S Du, Wei Hu, Ruslan Salakhutdinov, and Sanjeev Arora. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019.

[28] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[29] David Martin, Charless Fowlkes, Doron Tal, and Jitendra 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*, volume 2, pages 416–423. IEEE, 2001.

[30] Roey Mechrez, Itamar Talmi, Firas Shama, and Lihi Zelnik-Manor. Maintaining natural image statistics with the contextual loss. In *Asian Conference on Computer Vision*, pages 427–443. Springer, 2018.

[31] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018.

[32] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[33] Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. *arXiv preprint arXiv:1810.05148*, 2018.

[34] Edouard Oyallon, Eugene Belilovsky, and Sergey Zagoruyko. Scaling the scattering transform: Deep hybrid networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5618–5627, 2017.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[36] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[38] Ivan Ustyuzhaninov, Wieland Brendel, Leon A Gatys, and Matthias Bethge. What does it take to generate natural textures? In *ICLR (Poster)*, 2017.

[39] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010.

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.