Identifying and Mitigating Flaws of Deep Perceptual Similarity Metrics

Oskar Sjögren[1,2], Gustav Grund Pihlgren[1,2], Fredrik Sandin[1], and Marcus Liwicki[1]

[1]*Machine Learning Group* Luleå University of Technology, Sweden
[2]Equal contribution

Abstract

Measuring the similarity of images is a fundamental problem to computer vision for which no universal solution exists. While simple metrics such as the pixel-wise L2-norm have been shown to have significant flaws, they remain popular. One group of recent state-of-the-art metrics that mitigates some of those flaws are Deep Perceptual Similarity (DPS) metrics, where the similarity is evaluated as the distance in the deep features of neural networks. However, DPS metrics themselves have been less thoroughly examined for their benefits and, especially, their flaws. This work investigates the most common DPS metric, where deep features are compared by spatial position, along with metrics comparing the averaged and sorted deep features. The metrics are analyzed in-depth to understand the strengths and weaknesses of the metrics by using images designed specifically to challenge them. This work contributes with new insights into the flaws of DPS, and further suggests improvements to the metrics. An implementation of this work is available online.[1]

1 Introduction

Similarity metrics are a fundamental part of many machine learning processes. In computer vision, widely used metrics, such as the pixel-wise L2-norm, have been carefully studied and their benefits and flaws are well-known which lets users make an informed decision when using them.

Many improvements to pixel-wise metrics have been proposed, with a common goal being to mimic human perception with a so-called perceptual similarity. A recent approach is to utilize deep features learned by machine learning models for measuring perceptual similarity. This practice, called Deep Perceptual Similarity (DPS) measures the similarity of two images by comparing their respective activations in the deep layers of neural networks, instead of using the pixel values directly.

DPS metrics have outperformed previous models on perceptual similarity [31]. Additionally, such metrics have been used as part of the loss function for training models, which have achieved impressive results on a host of tasks. These tasks include, image generation [17], style transfer and super-resolution [12], object detection [19], and image segmentation [22].

While there are clear benefits of DPS, its flaws are not as well studied. Deep perceptual similarity is vulnerable to adversarial examples, which is expected from any method depending on deep networks. Existing methods for protecting from adversarial attacks such as ensembles may be utilized [13]. Additionally, adversarial examples are quite complex compared to the known flaws of other metrics.

This work aims to analyze if and how DPS can successfully handle the flaws of the pixel-wise L2-norm, and investigate if there are any similar unexplored flaws of DPS and how those may be mitigated. Additionally, several different DPS metrics are analyzed for flaws and then evaluated on the BAPPS dataset [31], to check if those flaws translate into performance on an actual dataset.

The investigation of DPS is performed by creating image pairs that are similar to each other compared to some reference images and checking

[1] `https://github.com/guspih/deep_perceptual_similarity_analysis/`

https://doi.org/10.7557/18.6795

in which cases the DPS metrics succeed or fail in identifying the image pairs as more similar than the reference. The feature maps of the CNNs used for calculating similarity are analyzed to gain insight as to what underlies the successes and failures.

2 Related Work

Pixel-wise metrics have long been known to be poor similarity metrics as they disregard high-level image structures [27, 28]. Instead many different perceptual similarity metrics have been proposed including Dynamic Partial Function [18], the Structural Similarity Index Measure [29], and Structural Texture Similarity [32]. Despite known flaws and suitable alternatives, pixel-wise metrics have consistently been used for image comparison within computer vision in general, and to calculate the loss for machine learning models specifically.

One powerful attribute of deep learning is that the deep features learned by the networks typically contain information useful for other tasks than the one the network was trained for. This attribute was used to great effect with the introduction of neural style transfer, where the content and style of images were compared using different sets of deep features within a neural network [7]. This practice of training models to minimize the difference between the activations of a deep network in order to get visually similar images is known as deep perceptual loss.

Deep perceptual loss has since its introduction been successfully applied to a large number of computer vision tasks such as improving the performance of variational autoencoders [9, 8, 2], Generative Adversarial Networks [17], Super-Resolution [6, 20], and style transfer [12]. The method has been proven effective at the task of perceptual similarity where it significantly outperformed previous methods [31]. This method of calculating perceptual similarity using the deep features of neural networks is referred to as deep perceptual similarity (DPS).

One potential problem with DPS is that it relies on deep neural networks, which are known to be vulnerable to adversarial examples. Adversarial examples are almost imperceptible perturbations to images or other input data that induce significant changes or errors to the prediction model [11]. While no perfect protection from adversarial examples is currently known, there is a wide array of defenses that can be used, including using ensembles [13].

Another paradigm for creating similarity metrics is to optimize a machine learning model for the task [23]. This has been applied to DPS with the LPIPS method, though it notably only performed marginally better than using methods that had only been pretrained [31]. Like with many other machine learning methods the results can be improved somewhat with the use of ensemble methods, though still comparable to pretrained models [13].

Where this work analyzes DPS through deep analysis of cases where it fails, another recent work investigates how different network architectures and pretraining procedures affect performance [16]. That work found, among other things, that better pretraining performance on ImageNet [5], does not necessarily lead to better perceptual similarity, It additionally showed that a good pretrained model can outperform models trained specifically for the similarity task.

As DPS metrics rely on the deep activations of neural networks, most commonly CNNs, analyzing these activations is inherently interesting. Many methods for such analysis exist and one of the most common is to visualize the feature maps of the CNNs [30], which is utilized in this work.

3 Deep Perceptual Similarity

Most uses of deep perceptual similarity and deep perceptual loss have directly compared the corresponding activations of the two images. This method, referred to as spatial DPS, is formalized as the distance measure between x and x_0 in Eq. 1, where f is a norm such as L1 or L2 and p is a convolutional feature extractor with extraction layers $l \in L$ each with C_l channels with height H_l, and width W_l.

$$d(x, x_0) = \sum_{l}^{L} \frac{1}{C_l H_l W_l} \sum_{c,h,w}^{C_l,H_l,W_l} f(p(x)_{lc}^{hw} - p(x_0)_{lc}^{hw}) \tag{1}$$

This work evaluates two additional methods of calculating deep perceptual similarity besides the spatial method. These two are the mean method tested in [16] and a sort method that is introduced in this work. The two methods are formalized in Eq. 2 Eq. 3 where $\overline{x}$ and $x^{\downarrow}$ are the average and descending reordering of x respectively.

$$d(x, x_0) = \sum_{l}^{L} \frac{1}{C_l} \sum_{c}^{C_l} f(\overline{p(x)_{lc}} - \overline{p(x_0)_{lc}}) \qquad (2)$$

$$d(x, x_0) = \sum_{l}^{L} \frac{1}{C_l} \sum_{c}^{C_l} f(p(x)^{\downarrow}_{lc} - p(x_0)^{\downarrow}_{lc}) \qquad (3)$$

Both of these methods ignore the spatial positions of the features. The mean method compares the average of the features in each channel and the sort method pairs the features of each channel with one another in such a way as to minimize the norm. In the sort method the norm is minimized for any convex function f, compared to any other ordering of the features. This follows from $x \prec y \rightarrow \sum f(x) \leq \sum f(y)$ and $a^{\downarrow} - b^{\downarrow} = a^{\downarrow} + (-b)^{\uparrow} \prec a + b$ [21]. The reasoning behind comparing average and sorted channels is that a strong activation in one channel often represents different concepts than a similar activation in another.

A problem with the mean and sort methods on their own is that humans would likely say that a lower translation is more similar to the original than a greater one. As such completely ignoring spatial position is not desirable. Thus, this work also investigates metrics that use the sum of the spatial method with one of the two non-spatial methods.

3.1 Experimental Setup

DPS relies on neural networks whose deep features contain useful information for image comparison. While networks can be trained specifically for the task, the most common use of DPS and deep perceptual loss is pretrained networks.

This work uses mostly the same feature extraction and comparison setup as [31]. The methods are analyzed and evaluated with the L2-norm as the comparison function (f) using three models (p) pretrained on the ImageNet dataset [5]. The architectures for the three models are SqueezeNet [10], AlexNet [15, 14], and VGG-16 [25]. The deep features are extracted from the same multiple layers for each network as in [31]. The features extracted in the original work were channel-wise unit-normalized, and this work analyzes and evaluates both using and ignoring this practice. However, for brevity, the analysis in Section 4 concerns only the case without unit-normalization and the use of unit-normalization is later discussed in Section 7.

4 Qualitative Analysis of DPS on Distortions

This work carries out a qualitative analysis of deep perceptual similarity metrics over images specifically designed to test for its strengths and potential flaws. The analysis is carried out by distorting images in ways for which DPS is previously known to work well or speculated to perform poorly. The similarity of the distorted image with the original is then compared to the similarities of a set of reference images and the original, where the reference images are intended to be notably less similar than the distortion. The feature maps at various layers of the DPS networks are then analyzed for each case to gain a deeper understanding of why the metric performed the way it did in this case. Such insight is then used to create further image pairs to test against. Finally, for one category of images, specific reference images were created for each image pair. These reference images, like the others, were created to be perceived by humans as less similar than the distorted versions but specific to that image pair. An aggregation of how well the different metrics identified the correct image to be more similar can be found in Table 1.

The images used in the tests are 96 × 96 pixels and have been designed and distorted by hand. The distortions tested are divided into four categories; color inversion, translation, rotation, and color stain. Seven reference images were created; mono-colored images of black, white, gray, red, green, and blue, as well as one with randomly colored pixels.

4.1 Black-and-White Color Inversion

Color inversion of black-and-white images is typically used as an example of when pixel-wise metrics break down since each pixel now has the opposite color. Despite this being used as an example of why pixel-wise metrics are worse than DPS metrics, there has been little investigation of how well DPS performs in these scenarios. For these reasons the first set of images created for analyzing DPS were color-inverted black-and-white patterns.

While pixel-wise metrics fail by definition on this category of images, all tested DPS metrics get almost perfect scores. Analysis of the feature maps reveals that many channels are activated by contrasts or higher-level structures like lines or shapes. These activations are often completely agnostic to inversion and identify the structures regardless of color. This makes the black-and-white inversion pairs almost exactly the same for many channels in the feature space, which leads to good performance.

4.2 Translation and Rotation

It is also clear from the feature maps that all activations are strongly spatially correlated to where those features appear in the input image, which can be seen in Fig. 2. This is obvious as CNN architectures in general are built around each activation depending only on a small region of the input or previous layer. This has been previously suggested as a potential flaw of spatial DPS [16].

To investigate whether this would have a significant impact on spatial DPS and whether other DPS metrics could handle these cases, the categories of translation and rotation have been tested. The translation images have a region containing much structure in otherwise plain images which have been distorted by translating that region. The rotation images are simply images that have been distorted by rotation in steps of 22.5 up to 90 degrees, as well as one rotated 180 degrees.

Both the pixel-wise metric and spatial DPS fail to identify any translated image as more similar than the reference images, while the other DPS metrics succeed in each case. For rotation, both pixel-wise and spatial DPS metrics fail on about the same amount of cases, slightly less than half, while the other DPS metrics almost succeed on each image pair. This clearly shows that the spatial DPS metric on its own is not suitable for these types of scenarios, while translation-invariant DPS metrics can handle them very well.

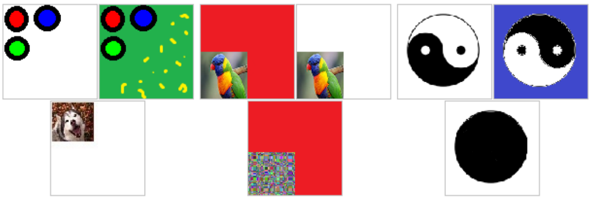

Figure 1: Image pairs from the color stain category (above) with their specific reference images (below).

4.3 Color Stain

Another revelation from feature map analysis is that many channels tend to activate strongly from specific colors, textures, or random noisy structures. This might be challenging for non-spatial methods as ignoring the spatial position of activations might lead to confusing noise for interesting structures. To test for this the color stain category is used.

The image pair for the color stain category consists of a plain image with a structurally interesting region, and a distorted version with a similar or same interesting region but the plain color is changed, and noisy features are added for some images. The color stain category uses specific reference images for each pair, instead of the ones used previously. These reference images have the same plain color as the non-distorted image, but their interesting region is significantly different compared to the distorted version. Examples of image pairs and their specific reference image are shown in Fig. 1.

For the color stain category, the pixel-wise metric again fails for each image pair. Notably, both the mean and spatial+mean DPS fails almost all image pairs. The remaining DPS metrics tested perform well, with spatial DPS being the best.

One specific image in this category is a white image with a red, green, and blue irregular circle in one corner. The distorted image retained the circles but the plain white background was colored a darker shade of green with random yellow stains. By observing the feature maps of these images it is clear that the color change and stains add significant activations to the otherwise sparse feature maps, especially in later layers. This is shown in Fig. 2, where the image and its distortion are dis-

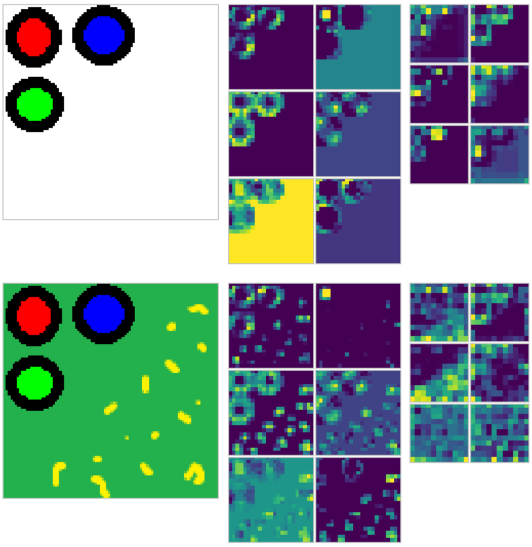

Figure 2: An image (top-left) and its color stain distorted version (bottom-left) with their respective feature maps from the second (middle) and fourth (right) ReLU layer.

played together with feature maps from the second and fourth SqueezeNet ReLU layer.

5 Evaluation

In order to investigate how the insights from the qualitative analysis translate to performance on a perceptual similarity dataset, the DPS metrics are evaluated on the BAPPS dataset, using the same procedure as in the original work [31]. This evaluation follows that of the pretrained networks in the original work which means f is the L2 norm and the features extracted from p have been channel-wise unit-normalized.

5.1 BAPPS

BAPPS is an image dataset consisting of 64×64 image patches sampled from the MIT-Adobe 5k [3], RAISE1k [4], DIV2K [1], Davis Middlebury [24], video deblurring [26], and ImageNet [5] datasets as well as a host of distortions of those same patches. The BAPPS dataset consists of two sets with different labels and intended use, Two Alternative Forced Choice (2AFC) and Just Noticeable Differences (JND).

2AFC consists of image patches and two distorted versions of each patch, as well as human annotations as to which distorted patch is most similar to the original. The task of 2AFC is to predict human similarity judgments. The 2AFC part ha six subdivisions defined by the type of distortions that are applied: (1) **Traditional** augmentations, outputs from (2) **CNN-based** autoencoders, (3) **superresolution**, (4) **frame interpolation**, (5) **video deblurring**, and (6) **colorization**.

JND consists of an image patch as well as a barely distorted version along with human annotations of whether they thought the two patches were the same after seeing them briefly. The task of JND is to make a model that gives a higher similarity to those samples that human annotators had difficulty telling apart.

6 Results

An aggregation of the outcome of the tests described in Section 4 is shown in Table 1. The performance is presented as the number of images, where the metric did not find any of the reference images to be more similar than the distorted version.

The results of the evaluation on the BAPPS dataset are shown in Table 2 for the 2AFC part as a whole, its subdivision, and for the JND part. The results for the evaluated metrics are presented along with human performance for reference. Note that, since the task is to estimate human perception, the human performance is also the maximum achievable.

7 Discussion

The purpose of this work has been (i) to evaluate if and how DPS metrics can handle the typical cases where pixel-wise metrics fail and (ii) to investigate whether similar flaws exist in current DPS implementations. All tested DPS metrics handle color inversion and the clear preference for structures in feature maps indicates that DPS metrics are well-suited to handle other similar cases. By far most common form of DPS metrics used is spatial DPS, which performs poorly on the rotation and translation test cases. While the non-spatial DPS metrics perform well on these weaknesses, they do

Table 1: Fraction of distorted images that were recognized as more similar than the reference images for the different metrics

Method	Network	Invert	Rotate	Translate	Color Stain
Pixel-Wise	-	0/11	17/30	0/5	0/5
Spatial	Squeeze	11/11	20/30	0/5	5/5
	AlexNet	11/11	11/30	0/5	3/5
	VGG-16	10/11	6/30	0/5	4/5
Sort	Squeeze	11/11	28/30	5/5	4/5
	AlexNet	11/11	30/30	5/5	2/5
	VGG-16	11/11	30/30	5/5	3/5
Mean	Squeeze	11/11	29/30	5/5	3/5
	AlexNet	11/11	28/30	5/5	2/5
	VGG-16	10/11	30/30	5/5	1/5
Spatial+Sort	Squeeze	11/11	22/30	0/5	5/5
	AlexNet	11/11	19/30	0/5	4/5
	VGG-16	11/11	21/30	1/5	4/5
Spatial+Mean	Squeeze	11/11	22/30	0/5	5/5
	AlexNet	11/11	15/30	0/5	2/5
	VGG-16	10/11	9/30	0/5	4/5

not perform as well as spatial metrics on the color stain category of tests. This is especially true for mean DPS which failed most of the color stain tests. The spatial and non-spatial combined metrics perform similarly to spatial DPS, indicating that perhaps combining metrics using unweighted summation gives a preference for spatial DPS. Though the combined metrics improved over spatial on rotation, indicating that there are some benefits to this strategy.

Analyzing the BAPPS scores for the different DPS metrics shows that spatial DPS, in general, performs worse than the other DPS metrics. This is especially true for the traditional augmentations which include operations such as rotation, translation, and skewing which indicates that the weaker is due to the flaws identified in this work.

Another notable result is that mean DPS, in general, performs best on BAPPS, even though it was most vulnerable to color stain distortions. However, both mean and sort DPS metrics perform similarly and are both better choices than spatial DPS. It is possible that color stain and related distortions are not so common to be a problem in a real-world scenario, or that the BAPPS dataset does not include many such cases.

7.1 The effects of unit-normalization

As mentioned in Subsection 3.1, the qualitative analysis described in Section 4 was also performed with channel-wise unit-normalization of the extracted features. This had three notable effects. First, the success rate in the rotate category rose for all DPS metrics. Second, the combined metrics are somewhat improved in the translate category. Likely due to normalizing making the spatial metrics lower which means the non-spatial metrics account for a larger fraction. Finally, using normalization made each DPS metric perform poorly in the color stain category.

On the BAPPS dataset unit-normalization has a small positive effect on performance. Likely augmentations similar to the color stain procedure are not common in the dataset.

8 Future Work

From the results and analysis presented in this work there are some notable directions of research to explore.

Both this and a prior work [16] has shown that spatial DPS does not perform as well as on perceptual similarity tasks as other implementations of DPS. One future possibility is to investigate if this translates to related field such as deep perceptual loss and content-based image retrieval. If it does, simply changing the way perceptual loss is calculated could improve the results on many different tasks.

While most DPS metrics outperform previous perceptual similarity metrics, the discrepancy in performance of DPS metrics indicates that exploring how to calculate DPS metrics is an open problem. For example, a DPS metric that make use of both spatial and non-spatial comparisons could perhaps gain the benefit of both. Additionally, the upsides and downsides of unit-normalization remain inconclusive.

Table 2: Performance of the evaluated DPS metrics on the BAPPS validation set (best values in bold)

		Distortions			Real Algorithms					All	JND
Method	Network	Traditional	CNN-based	All	Super-res	Video Deblur	Colorization	Frame Interp	All	All	JND
Human	-	80.8	84.4	82.6	73.4	67.1	68.8	68.6	69.5	73.9	-
Spatial	Squeeze	73.3	82.6	78.0	70.1	60.1	63.6	62.0	64.0	68.6	60.2
	AlexNet	70.6	**83.1**	76.8	71.7	60.7	65.0	62.7	65.0	68.9	57.6
	VGG-16	70.1	81.3	75.7	69.0	59.0	60.2	62.1	62.6	67.0	59.1
Mean	Squeeze	77.1	82.3	79.7	69.9	60.0	65.2	63.1	64.5	69.5	63.6
	AlexNet	73.9	82.8	78.4	71.4	60.7	65.5	63.5	**65.3**	**69.6**	60.2
	VGG-16	77.9	81.8	**79.8**	68.9	59.5	64.0	63.0	63.8	69.2	**65.2**
Sort	Squeeze	76.8	82.0	79.4	69.8	60.1	64.6	61.9	64.1	69.2	62.0
	AlexNet	73.3	82.8	78.0	71.1	60.6	64.6	62.6	64.7	69.2	58.5
	VGG-16	**78.1**	81.5	**79.8**	68.1	59.2	62.7	61.5	62.9	68.5	64.8
Spatial+Mean	Squeeze	75.0	82.5	78.8	69.9	60.1	64.5	62.1	64.2	69.0	61.5
	AlexNet	71.8	83.0	77.4	71.6	60.7	65.5	62.7	65.1	69.2	58.5
	VGG-16	73.4	81.9	77.7	69.3	59.4	64.5	62.5	63.9	68.2	61.0
Spatial+Sort	Squeeze	75.5	82.5	79.0	70.0	60.1	64.4	61.9	64.1	69.1	61.2
	AlexNet	72.2	**83.1**	77.7	71.3	60.6	64.9	62.8	64.9	**69.2**	58.5
	VGG-16	74.9	81.9	78.4	69.4	59.4	62.3	62.1	63.3	68.4	61.9

References

[1] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1122–1131, 2017. doi: 10.1109/CVPRW.2017.150.

[2] S. Bhardwaj, I. Fischer, J. Ballé, and T. Chinen. An unsupervised information-theoretic perceptual quality metric. In *Advances in Neural Information Processing Systems*, volume 33, pages 13–24. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/00482b9bed15a272730fcb590ffebddd-Paper.pdf`.

[3] V. Bychkovsky, S. Paris, E. Chan, and F. Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *CVPR 2011*, pages 97–104, 2011. doi: 10.1109/CVPR.2011.5995332.

[4] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato. Raise: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM multimedia systems conference*, pages 219–224, 2015. doi: 10.1145/2713168.2713194.

[5] 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*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[6] 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*, 38(2):295–307, 2016. doi: 10.1109/TPAMI.2015.2439281.

[7] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*

(CVPR), pages 2414–2423, 2016. doi: 10.1109/CVPR.2016.265.

[8] G. Grund Pihlgren, F. Sandin, and M. Liwicki. Pretraining image encoders without reconstruction via feature prediction loss. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4105–4111, 2021. doi: 10.1109/ICPR48806.2021.9412239.

[9] X. Hou, L. Shen, K. Sun, and G. Qiu. Deep feature consistent variational autoencoder. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1133–1141, 2017. doi: 10.1109/WACV.2017.131.

[10] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. doi: 10.48550/arXiv.1602.07360.

[11] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf`.

[12] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. doi: 10.1007/978-3-319-46475-6_43.

[13] M. Kettunen, E. Härkönen, and J. Lehtinen. E-lpips: robust perceptual image similarity via random transformation ensembles. *arXiv preprint arXiv:1906.03973*, 2019. doi: 10.48550/arXiv.1906.03973.

[14] A. Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014. doi: 10.48550/arXiv.1404.5997.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[16] M. Kumar, N. Houlsby, N. Kalchbrenner, and E. D. Cubuk. Do better imagenet classifiers assess perceptual similarity better? *Transactions of Machine Learning Research*, 2022. URL `https://openreview.net/forum?id=qrGKGZZvH0`.

[17] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1558–1566, New York, New York, USA, June 2016. PMLR.

[18] B. Li, E. Chang, and Y. Wu. Discovery of a perceptual distance function for measuring image similarity. *Multimedia systems*, 8(6):512–522, 2003. doi: 10.1007/s00530-002-0069-9.

[19] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan. Perceptual generative adversarial networks for small object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1951–1959, 2017. doi: 10.1109/CVPR.2017.211.

[20] A. Lucas, S. López-Tapia, R. Molina, and A. K. Katsaggelos. Generative adversarial networks and perceptual losses for video super-resolution. *IEEE Transactions on Image Processing*, 28(7):3312–3327, 2019. doi: 10.1109/TIP.2019.2895768.

[21] A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, 2 edition, 2011. doi: 10.1007/978-0-387-68276-1.

[22] A. Mosinska, P. Marquez-Neila, M. Kozinski, and P. Fua. Beyond the pixel-wise loss for topology-aware delineation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3136–3145, 2018. doi: 10.1109/CVPR.2018.00331.

[23] F. Ricci and P. Avesani. Learning a local similarity metric for case-based reasoning. In *Case-Based Reasoning Research and Development*, pages 301–312, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg. ISBN 978-3-540-48446-2. doi: 10.1007/3-540-60598-3_27.

[24] D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, pages 131–140, 2001. doi: 10.1109/SMBV.2001.988771.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. doi: 10.48550/arXiv.1409.1556.

[26] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang. Deep video deblurring for hand-held cameras. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 237–246, 2017. doi: 10.1109/CVPR.2017.33.

[27] C. C. Taylor. Measures of similarity between two images. *Lecture Notes-Monograph Series*, 20:382–391, 1991. ISSN 07492170. URL http://www.jstor.org/stable/4355717.

[28] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009. doi: 10.1109/MSP.2008.930649.

[29] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

[30] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1. doi: 10.1007/978-3-319-10590-1_53.

[31] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

[32] X. Zhao, M. G. Reyes, T. N. Pappas, and D. L. Neuhoff. Structural texture similarity metrics for retrieval applications. In *2008 15th IEEE International Conference on Image Processing*, pages 1196–1199, 2008. doi: 10.1109/ICIP.2008.4711975.