Deep Perceptual Similarity is Adaptable to Ambiguous Contexts

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

[1]Dept. of Computing Science, Umeå University
[2]EISLAB Machine Learning, Luleå University of Technology
gustav.pihlgren@umu.se, fredrik.sandin@ltu.se, marcus.liwicki@ltu.se

Abstract

This work examines the adaptability of Deep Perceptual Similarity (DPS) metrics to context beyond those that align with average human perception and contexts in which the standard metrics have been shown to perform well. Prior works have shown that DPS metrics are good at estimating human perception of similarity, so-called perceptual similarity. However, it remains unknown whether such metrics can be adapted to other contexts.

In this work, DPS metrics are evaluated for their adaptability to different contradictory similarity contexts. Such contexts are created by randomly ranking six image distortions. Metrics are adapted to consider distortions more or less disruptive to similarity depending on their place in the random rankings. This is done by training pretrained CNNs to measure similarity according to given contexts. The adapted metrics are also evaluated on a perceptual similarity dataset to evaluate whether adapting to a ranking affects their prior performance.

The findings show that DPS metrics can be adapted with high performance. While the adapted metrics have difficulties with the same contexts as baselines, performance is improved in 99% of cases. Finally, it is shown that the adaption is not significantly detrimental to prior performance on perceptual similarity.

The implementation of this work is available online [1].

1 Introduction

Image similarity metrics are used in many tasks and methods. Research on image similarity has focused on so-called perceptual similarity, where the goal is to approximate human (or animal) perception of similarity. Perceptual similarity metrics can be directly applied to tasks such as image retrieval [1] and image quality assessment [2].

Recently a method called deep perceptual similarity (DPS) has achieved close to human performance on perceptual similarity [3, 4]. DPS metrics compare the difference between the deep features (activations) of a neural network, called the loss network when the input is one image compared to another. This approach has been used to calculate deep perceptual loss (DPL) for computer vision models. DPL has been successfully applied to image generation [5], style-transfer [6], and super-resolution [7], image segmentation [8], depth prediction [9], and more.

However, perceptual similarity has long been known to be an ambiguous concept [10], with the perception of similarity varying between populations and even within individuals as the context or focus changes. Additionally, it has been shown that there is no strong correlation between loss network performance on DPS and DPL [11]. This introduces further ambiguity as performance can be measured either by agreement with humans or downstream performance. The context of the image also affects how similarity should be measured. For example, blurring a medical image would significantly hamper the similarity of the samples, while the same operation could leave a photo of a cat recognizable [12].

Ambiguity raises the question of whether perceptual similarity metrics can adapt to varying contexts. Some rule-based metrics have hyperparameters that can be altered to fit the metric to particular circumstances, though the hyperparameters are typically limited in how they can alter the metrics [13]. DPS metrics, on the other hand, are based on neural networks that could theoretically be retrained to suit a given context.

Training a neural network for each context would be resource-intensive. For this reason, DPS is commonly implemented with pretrained networks, and the most common uses of DPL utilize pretrained networks as well. These networks are typically pretrained using ImageNet [14], an image dataset the size of which makes it computationally intensive to train on. Such ImageNet pretrained DPS metrics are the baseline for DPS metrics in this work.

As training on ImageNet is computationally expensive it might be expensive to retrain metrics for each given context. However, it is possible that no retraining is needed. Networks trained on ImageNet learn a large number of features that might be useful for many different definitions of similarity. So rather than retraining the network itself, a layer of scalars could be learned to balance the relevant features for a given circumstance.

*Corresponding Author.
[1]https://github.com/LTU-Machine-Learning/Analysis-of-Deep-Perceptual-Loss-Networks

Proceedings of the 5th Northern Lights Deep Learning Conference (NLDL), PMLR 233, 2024.

However, it is unknown whether DPS metrics based on ImageNet pretrained CNNs can be trained to adapt to certain contexts, especially those where the baseline pretrained metrics perform poorly. For example, CNN architectures are known to have flaws that make them vulnerable to certain distortions of the input [15]. Additionally, ImageNet pretrained CNNs have been shown to be biased toward the texture of the image over other structures [16].

Zhang et al. [3] showed that training positive scalars to weigh the different extracted feature maps could be used to improve the performance of perceptual similarity for the specific image-distortion distribution that the networks were trained on. However, performance decreased for image distortions outside the training set. Additionally, this was only examined for improving performance on agreement with human judgments of similarity which the ImageNet pretrained DPS metrics already performed well on. Whether DPS metrics can be adapted to contexts where the baseline metrics perform poorly remains unknown. Knowing if this is possible could unlock further use cases on domains where the average human perception is not as applicable or even detrimental.

This work investigates whether DPS metrics can be adapted to a specific context by training positive scalars that weigh the extracted features. The limitation to positive scalars may hamper the performance of the adapted metrics but allows analysis of whether the features learned by the pretrained networks are sufficient for adaption. The contexts used are created by randomly ranking six common image distortions. For a given ranking a distorted image is considered more similar to the original image than another distorted image if the distortion type applied to the first is earlier in the ranking. 20 contexts of random rankings are used for evaluation which provides a spectrum of contexts where the baseline DPS metrics either perform

The DPS metrics that are evaluated consist of combinations of three ImageNet pretrained CNNs and five methods of comparing the extracted features. The metrics are adapted to each context by training scalars in the same way as was done by *Zhang* et al. [3], to recognize some distortions as more similar than others. The metrics are adapted on images from the SVHN dataset [17] and evaluated by how well their similarity scores align with the given ranking on images from the test sets of SVHN and STL-10 [18]. Additionally, the adapted metrics are evaluated on the BAPPS dataset [3] to evaluate how the adaption affects their performance on known human judgments. For reference, pretrained baseline metrics without adaption are also evaluated.

The results show that metrics can be adapted to the given contexts and outperform baseline metrics.

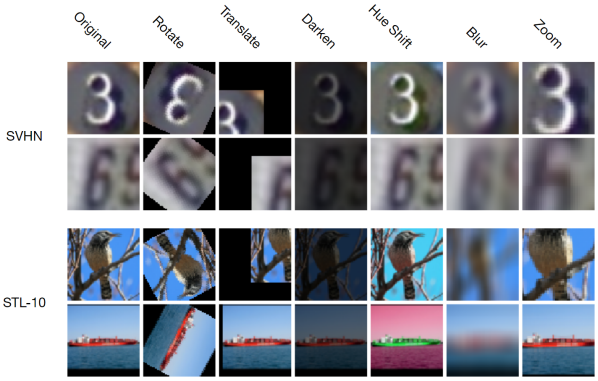

Figure 1. Example images from the test sets of SVHN and STL-10 with the six distortion types used in this work applied.

The performance of baseline and adapted metrics on the same ranking are shown to be correlated, meaning they perform well on the same rankings. This also means that while adaption can improve performance for contexts where baseline models perform poorly, the performance of adapted metrics is limited by the features of the baseline metrics being better suited for some contexts. Additionally, the adaption training has a slim detrimental effect on the performance on BAPPS.

The work also discovers a potential flaw in the established training procedure for DPS metrics that are used in this work. Improvements inspired by the field of contrastive learning in which different training methods adapted to specific domains are being explored [12] are discussed.

2 Datasets

Three datasets are used in this work, SVHN [17], STL-10 [18], and BAPPS [3].

SVHN and STL-10 are image classification datasets, but only the images are used in this work. The SVHN training images are used to train metrics to adapt to a given ranking of distortions and the test images of SVHN and STL-10 are both used for testing the adaption. The two datasets are used since they are drawn from different distributions, which tests whether adaptions learned on one image distribution generalize to another. The difference is shown in Fig. 1, which shows samples from both.

BAPPS is a perceptual similarity dataset used in this work to evaluate how adaption affects the metrics performance on human perception of similarity. BAPPS is split into two parts, Two-Alternative Forced-Choice (2AFC) and Just Noticeable Differences (JND). 2AFC consists of triplets of an image (x) and two distorted versions (x_0 and x_1), labeled by the fraction of human judges (J) that considered x_0 to be more similar to x. JND consists of slightly distorted image pairs labeled by the fraction of hu-

Table 1. Loss networks and feature extraction layers.

Architecture	Feature Extraction Layer
AlexNet [19]	1st, 2nd, 3rd, 4th, and 5th ReLU
SqueezeNet 1.1 [20]	1st ReLU 2nd, 4th, 5th, 6th, 7th and 8th Fire
VGG-16 [21]	2nd, 4th, 7th, 10th, and 13th ReLU

man judges that considered the pair to be the same image.

Table 2. Distortions and the intervals from which their parameters are randomly chosen.

Distortion	Parameters and intervals
Rotating	30 to 330 degrees
Translating	−0.5 to 0.5 of image size
Lowering brightness	0.1 to 0.5 of original brightness
Shifting hue	−0.5 to 0.5 hue factor
Gaussian blurring	11 to 21 kernel size 4 to 10 std. dev. for kernel values
Zooming in	1.1 to 2 scale of zoom

3 Methodology

This work investigates whether the deep features of ImageNet [14] pretrained CNNs contain the necessary information to adapt to different definitions of similarity and if this adaption can be achieved by learning scalars of the features for each definition. To do this an altered version of the experiments by *Zhang* et al. [3] are used.

The experiments test many combinations of loss networks, feature comparison methods, and training procedures. Three loss networks pretrained on ImageNet [14] with different architectures are used for extracting features at different layers. Five methods for comparing the extracted features are used to create metrics for each loss network. The metrics are evaluated both as baseline metrics without extra training and as adapted metrics trained on a specific ranking of distortions. Each combination is trained and tested on 20 different random rankings of distortions. Each adapted metric being trained four times to evaluate variance in training. All of these parts are detailed in the following subsection.

3.1 Loss Networks

This work uses the same three loss networks as *Zhang* et al. [3]. They are AlexNet [19], SqueezeNet 1.1 [20], and VGG-16 [21] pretrained on the ImageNet [14] dataset. The specific implementation of each architecture and the trained model parameters were taken from the Torchvision [22] framework version 0.11.3. The features were extracted from layers throughout the convolutional parts of the models as detailed in Table 1.

3.2 Similarity Calculations

The similarity between two images is calculated by using them each as input to the same loss network and then using the difference between the extracted deep features of each image as a distance metric. This work uses the spatial, mean, and sort comparison methods used by *Sjögren* et al. [23] for feature comparison. The methods are detailed in Eq. 1 to 3 below, where z^l_x are the activations in layer l from a loss network with input x and extraction layers $l \in L$. $\overline{z}$ and $z^{\downarrow}$ are the average and descending sorting of the channels in z respectively. w_l are the scalars for the features of layer l, which are set to 1 in the baseline cases and adapted to be positive values during adaption training. In addition, two combinations are used consisting of the sum of the spatial and mean metrics ($d_{spatial+mean} = d_{spatial} + d_{mean}$), as well as spatial and sort ($d_{spatial+sort} = d_{spatial} + d_{sort}$).

$$d_{spatial}(x, x_0) = \sum_{l \in L} \frac{1}{C_l H_l W_l} ||w_l \odot (z^l_x - z^l_{x_0})||_2^2 \quad (1)$$

$$d_{mean}(x, x_0) = \sum_{l \in L} \frac{1}{C_l} ||w_l \odot (\overline{z}^l_x - \overline{z}^l_{x_0})||_2^2 \quad (2)$$

$$d_{sort}(x, x_0) = \sum_{l \in L} \frac{1}{C_l} ||w_l \odot (z^{l\downarrow}_x - z^{l\downarrow}_{x_0})||_2^2 \quad (3)$$

3.3 Distortions

To train and evaluate metrics for their ability to adapt to varying definitions of similarity, six distortions taken from commonly applied image augmentation procedures [24] are used. The distortions are rotating, translating, lowering brightness, shifting hue, Gaussian blurring, and zooming in. The distortions are implemented using the Torchvision [22] framework and are applied with parameters chosen randomly from the intervals shown in Table 2. Fig. 1 shows random applications of the six distortion types to images from test sets of SVHN and STL-10.

3.4 Adaption Training

Metrics are adapted to each ranking of the distortions by training the scalars (w) using the images from the SVHN training set. For each image, a 2AFC triplet is created with the original image and two versions each distorted by one of the six distortions chosen at random. The triplet is labeled 0 if the first distortion is earlier in the ranking and 1 otherwise. The metric is then used to calculate the similarity scores between the original image and the two distortions.

During training, loss is given by the Binary Cross-Entropy (BCE) between the labels and predictions

by an auxiliary CNN that classifies which image is more similar using the similarity scores as input. This is the same as the 2AFC training implemented by *Zhang* et al. [3], and exact details can be found in that work. An additional loss $\mathcal{L}_{sync}$ is used until the validation 2AFC score is higher than random for one epoch (0.5). The loss is detailed in Eq. 4 where d is the metric being trained, x is the original image, x_0 and x_1 are the distorted versions, J is 0 if the distortion of x_0 is earlier in the ranking and 1 otherwise, and σ is the sigmoid function.

$$\mathcal{L}_{sync}(x, x_0, x_1, J) = \\ 10 \cdot \max(0, \text{BCE}(\sigma(d(x, x_0) - d(x, x_1)), J)) \tag{4}$$

Training is performed for 10 epochs with validation using 20% of the training data. During the last 5 epochs, the learning rate decays linearly towards 0. Each adapted metric is trained four times in order to measure the variance of training.

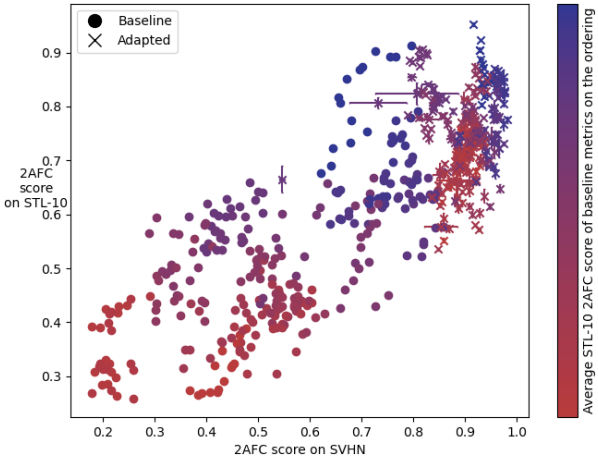

Figure 2. The 2AFC score of baseline (•) and adapted (×) metrics on SVHN and STL-10. The color indicates which ranking the metric was evaluated on, going from the ranking with the lowest STL-10 2AFC score for baselines (red) to the highest (blue).

3.5 Evaluation

For each ranking, each baseline metric, and each metric adapted to that ranking four performance scores were gathered. The first two were gathered by taking the test set images in SVHN and STL-10 and creating 2AFC triplets consisting of the image and two versions of it distorted by two different randomly chosen distortions. The 2AFC score of the metrics was calculated by whether they consider the version whose distortion is earlier in the ranking to be more similar. The two remaining performance scores are the 2AFC and JND scores for the respective parts of the BAPPS dataset. The JND score is the mean average precision between the image pairs when ordered by how many human annotators judged them to be the same and the order of similarity produced by the metric. The calculation of the 2AFC score for a single sample is detailed in Eq. 5 for distance metric d, an image x, distorted versions x_0 and x_1, and the fraction J of judgments that consider x_1 more similar to x than x_0. In the SVHN and STL-10 evaluations, J is 0 if the distortion used for x_0 is earlier in the ranking and 1 otherwise. The final 2AFC score is the average score for each sample.

$$\text{2AFC}(x, x_0, x_1, J) = \begin{cases} J, & \text{if } d(x, x_1) < d(x, x_0) \\ 1 - J, & \text{otherwise} \end{cases} \tag{5}$$

4 Results and Analysis

The 2AFC scores for measuring similarity according to the rankings on SVHN and STL-10 is shown in Fig. 2. The figure shows the score for each combination of ranking, loss network, comparison method, and whether the metric is baseline (•) or adapted (×). The four versions of each adapted metric are shown as an average and standard deviation. The points are colored by ranking, with red indicating rankings where baseline models had low STL-10 2AFC scores and blue high scores.

The adapted metrics significantly outperform the baseline on SVHN and mostly outperform them on STL-10 as well. Out of 300 adapted metrics, only 2 do not outperform their baseline counterparts on average, and in both cases, the performance difference is ~ 0.01. The disparity likely comes from the metrics being adapted specifically to SVHN which contain quite different images to STL-10.

The baseline and adapted metrics also perform better on the same rankings. The Spearman correlation between the baseline and adapted average performances for each ranking are 0.66 and 0.72 for SVHN and STL-10 respectively. This would likely not have been the case if the adapted metrics had been allowed to learn negative scalars since a metric that performs worse than random can be improved by simply inverting its predictions. This shows that the features present in the baseline metrics are sufficient for adapting with decent performance to each evaluated context. However, the adaptions do not generalize well to images from another distribution than they were trained on. Additionally, the correlation suggests that the features are better for certain contexts than others.

Fig. 3 shows the average performance on all rankings on STL-10 images for the different loss networks and comparison methods for baseline (lower bars) and adapted metrics (upper bars). It is clear that the adapting metrics provide a significant advantage on the average ranking. The figure also indicates that the specific loss network and comparison method does not significantly impact results, at least among

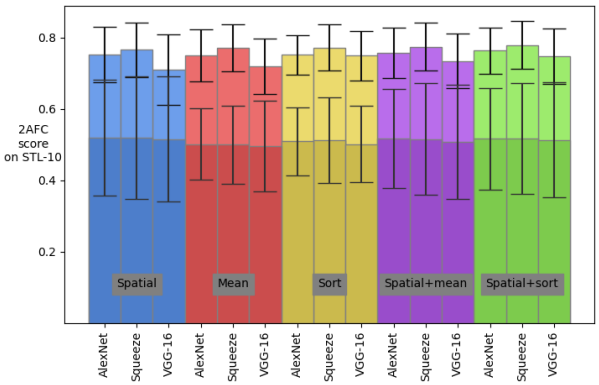

Figure 3. The average 2AFC score on STL-10 on all rankings for each loss network and comparison method. The lower bars are the baseline metrics and the upper their adapted counterparts.

Table 3. Average performance of metrics with AlexNet models and the spatial comparison method.

	BAPPS		SVHN	STL-10
Method	2AFC	JND	2AFC	2AFC
Baseline	0.6895	0.5757	0.5312	0.5197
Adapted Scalars	0.6817	0.5663	0.8916	0.7517
Adapted Net	0.6644	0.5552	0.9921	0.7961

the networks and methods evaluated in this work. The equivalent performance does not stem from different networks and methods performing well on different contexts. The average Pearson correlation coefficient for the 2AFC on STL-10 on all rankings between each combination of networks and methods is 0.79, which shows that the metrics on average perform well on the same rankings.

The adaption training is slightly detrimental to the metrics' performance on both the 2AFC and the JND parts of the BAPPS dataset. Both scores lower by ~ 0.01 on average for the adapted metrics compared to their baseline counterparts. It is interesting that the average baseline outperforms the adapted metrics on BAPPS on all rankings. This suggests that adaption training the entire loss network might be even more detrimental than the simple scalars learned in this work as it would have the potential to completely alter the features used.

To test whether adapting the entire loss network would be even more detrimental to the BAPPS scores an additional smaller run of experiments was conducted where DPS metrics with AlexNet architecture and spatial comparison were adapted by fine-tuning the parameters of the AlexNet model, in addition to training the scalars. Table 3 shows the results from this trial.

Adapting the entire network leads to better performance on the adapted context, but decreased performance on BAPPS. A significant part of the performance boost on SVHN and STL-10 can be attributed to allowing the inversion of features that the positive scalars could not perform. This is made clear by the lower correlation between which rankings the baselines perform well at compared to which the adapted scalars and adapted network metrics perform well on. The Spearman correlation between the baseline and scalar adaption is 0.70 and it is 0.35 between the baseline and network adaption.

5 Discussion

The results show that DPS metrics can be adapted through learning positive scalars for the extracted features to a context given by which distortions should be considered more similar. Additionally, the adapted metrics generalize to images from another distribution. This suggests that the features of ImageNet pretrained CNNs can be made suitable for similarity measurement in different contexts by weighing them differently. However, the performance is not great for all rankings, especially on the out-of-distribution images. Some of the lackluster performance can be attributed to forcing positive scalars. Even with negative scalars allowed, the rankings on which the baseline metrics have close to random performance would likely still be difficult. Whether this is applicable beyond the proof-of-concept contexts that have been examined would require further evaluation.

Adaption to a context is shown not to be significantly detrimental to the performance on BAPPS. This is desirable because it allows adapting metrics to specific contexts without significant risk of losing other desirable properties. For example, a metric could be adapted to deal with some invariances in the data without having to simultaneously train it on the original data. However, adapting metrics by fine-tuning the loss network was further detrimental to performance which could lead to collapse issues seen in continuously updated models [25]. Another approach might be to integrate the adaption with the pretraining of the loss network. *Kumar* et al. [4] have shown that the pretraining procedure significantly impacts on perceptual similarity performance. Including the adaption context among the other pretraining data or pretraining on a dataset specific to the domain could improve performance. Though this would likely be resource-intensive.

That performance is equivalent between different comparison methods is surprising. Previous work has shown that spatial DPS metrics struggle with translation and rotation [23], both of which are included in the rankings.

One potential answer is that the metrics might have learned to classify the distortions rather than measure their impact on similarity. For example, the translation and rotation operations used in this work color the missing pixels black. The metrics

might learn to discover if there is black in the image and then weigh that according to where the two distortions show up in the ranking. Such a quirk would be discoverable by all comparison methods, making their differences less impactful. It is also noted that the metrics might not be adapting to measure the similarity of images but rather to distinguish different distortion types from each other. If this is the case, it likely arises from only training on triplets from the same image, meaning that being able to classify distortions is enough to achieve high accuracy. This could be solved by taking inspiration from contrastive learning, which forms negative pairs from different images. Interestingly, this issue would also be present in BAPPS training that has been conducted by prior works [3, 4, 26]. The same solution would be applicable in this case as well, which could likely improve performance on the dataset even further. Such improved training may also prove efficient at solving the generalization issues found in this work, as well as those found by Zhang et al. [3].

6 Future Work

There are many promising future directions in ambiguity and adaptability. The adaption method explored in this work could be applied to more realistic scenarios. Medical images and non-RGB sensors where the perception of similarity might differ from natural images are interesting cases. For these domains and other applications, the question of how to get the data for adaption training is raised. Perhaps the use of distortions is applicable, where the rankings are defined by experts in the field.

Similar questions are being explored in the field of contrastive learning. In contrastive learning, distortions are used to learn features that are similar even as the image is distorted. However, it has been noted that the distortions that work well for natural images from ImageNet, do not generalize to other domains [12]. Instead, information in the data can be used to determine which distortions to use [12]. For perceptual similarity, it is not desirable to learn that two distortions of an image are the same, but similar approaches might be applicable.

DPL is another domain where adapting metrics to a specific context is useful. Loss functions can be evaluated by how well they work to train models for a given task. Adapting loss networks to calculate DPL for a given task might be beneficial. Which loss function works best might also change as training progresses, so making loss networks that target different stages of training is another idea.

Tasks where similar images have pixel-level differences are likely well suited to explore the impact of ambiguity and adaptability. The field of image retrieval has already acknowledged the issues that come with ambiguity [27, 28]. For example, when performing an image reverse search the desired result can vary heavily even for the same input image. One user might want to find images with the same composition while another wants to find scenes containing similar objects. In these scenarios having several metrics based on different contexts would be useful.

7 Conclusion

This work explores whether DPS metrics can be adapted to contexts where similarity does not adhere to the average human perception and where baseline metrics perform poorly. The contributions of this work consist of a proof-of-concept evaluation and analysis of the results. The proof-of-concept evaluation which shows that adaption improves performance on such contexts in 99% of cases, but that the performance of adapted metrics is limited by the performance of the baseline metrics. It is also shown that fine-tuning leads to better adaption but at the cost of prior performance, which may be a sign of overfitting. The analysis reveals a potential flaw in the training procedure commonly used for DPS metrics. Further work is proposed to address this flaw using training methods inspired by the field of contrastive learning.

References

[1] W. Hsu, S. Chua, and H. Pung. "An integrated color-spatial approach to content-based image retrieval". In: *Proceedings of the third ACM international conference on Multimedia*. 1995, pp. 305–313.

[2] R. Kazmierczak, G. Franchi, N. Belkhir, A. Manzanera, and D. Filliat. "A study of deep perceptual metrics for image quality assessment". In: arXiv preprint, 2022. DOI: 10.48550/ARXIV.2202.08692.

[3] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[4] M. Kumar, N. Houlsby, N. Kalchbrenner, and E. D. Cubuk. "Do better ImageNet classifiers assess perceptual similarity better?" In: *Transactions on Machine Learning Research* (2022). ISSN: 2835-8856.

[5] 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*. Vol. 48. Proceedings of Machine Learning Research. PMLR, June 2016, pp. 1558–1566.

[6] 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 (CVPR)*. June 2016.

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

[8] A. Mosinska, P. Marquez-Neila, M. Koziński, and P. Fua. "Beyond the pixel-wise loss for topology-aware delineation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3136–3145.

[9] X. Liu, H. Gao, and X. Ma. "Perceptual losses for self-supervised depth estimation". In: *Journal of Physics: Conference Series* 1952.2 (June 2021), p. 022040. DOI: 10.1088/1742-6596/1952/2/022040.

[10] L. B. Smith and D. Heise. "Perceptual similarity and conceptual structure". In: *Advances in psychology*. Vol. 93. Elsevier, 1992, pp. 233–272.

[11] G. G. Pihlgren, K. Nikolaidou, P. C. Chhipa, N. Abid, R. Saini, F. Sandin, and M. Liwicki. "A Systematic Performance Analysis of Deep Perceptual Loss Networks Breaks Transfer Learning Conventions". In: arXiv preprint, 2023. DOI: 10.48550/ARXIV.2302.04032.

[12] P. Chandra Chhipa. "Self-supervised Representation Learning for Visual Domains Beyond Natural Scenes". Licentiate Thesis. Luleå tekniska universitet, 2023.

[13] A. Eskicioglu and P. Fisher. "Image quality measures and their performance". In: *IEEE Transactions on Communications* 43.12 (1995), pp. 2959–2965. DOI: 10.1109/26.477498.

[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*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[15] A. Azulay and Y. Weiss. "Why do deep convolutional networks generalize so poorly to small image transformations?" In: *Journal of Machine Learning Research* 20.184 (2019), pp. 1–25.

[16] 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: *7th International Conference on Learning Representations ICLR*. 2019.

[17] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. "Reading digits in natural images with unsupervised feature learning". In: *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*. 2011.

[18] A. Coates, A. Ng, and H. Lee. "An analysis of single-layer networks in unsupervised feature learning". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. 2011, pp. 215–223.

[19] A. Krizhevsky. "One weird trick for parallelizing convolutional neural networks". In: arXiv preprint, 2014. DOI: 10.48550/ARXIV.1404.5997.

[20] 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 <0.5MB model size". In: arXiv preprint, 2016. DOI: 10.48550/ARXIV.1602.07360.

[21] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *3rd International Conference on Learning Representations ICLR*. 2015.

[22] S. Marcel and Y. Rodriguez. "Torchvision the Machine-Vision Package of Torch". In: *Proceedings of the 18th ACM International Conference on Multimedia*. MM '10. Association for Computing Machinery, 2010, pp. 1485–1488. ISBN: 9781605589336. DOI: 10.1145/1873951.1874254.

[23] O. Sjögren, G. G. Pihlgren, F. Sandin, and M. Liwicki. "Identifying and Mitigating Flaws of Deep Perceptual Similarity Metrics". In: *Proceedings of the Northern Lights Deep Learning Workshop 2023*. 2023. DOI: 10.7557/18.6795.

[24] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. "A Survey on Contrastive Self-Supervised Learning". In: *Technologies* 9.1 (2021). ISSN: 2227-7080. DOI: 10.3390/technologies9010002.

[25] R. M. French. "Catastrophic forgetting in connectionist networks". In: *Trends in Cognitive Sciences* 3.4 (1999), pp. 128–135. ISSN: 1364-6613. DOI: https://doi.org/10.1016/S1364-6613(99)01294-2.

[26] M. Kettunen, E. Härkönen, and J. Lehtinen. "E-LPIPS: Robust Perceptual Image Similarity via Random Transformation Ensembles". In: arXiv preprint, 2019. DOI: 10.48550/ARXIV.1906.03973.

[27] S. Saha and S. Sen. "Agent Based Framework for Content Based Image Retrieval". In: *Papers from the 2004 AAAI Spring Symposium*. AAAI Press, 2004.

[28] L. Rossetto, C. Tănase, and H. Schuldt. "Dealing with Ambiguous Queries in Multimodal Video Retrieval". In: *MultiMedia Modeling*. Cham: Springer International Publishing, 2016, pp. 898–909. ISBN: 978-3-319-27671-7.