# The Level Weighted Structural Similarity Loss: A Step Away from MSE

**Yingjing Lu**
Carnegie Mellon University
yingjinl@andrew.cmu.edu

## Abstract

The Mean Square Error (MSE) has shown its strength when applied in deep generative models such as Auto-Encoders to model reconstruction loss. However, in image domain especially, the limitation of MSE is obvious: it assumes pixel independence and ignores spatial relationships of samples. This contradicts most architectures of Auto-Encoders which use convolutional layers to extract spatial dependent features. We base on the structural similarity metric (SSIM) and propose a novel level weighted structural similarity (LWSSIM) loss for convolutional Auto-Encoders. Experiments on common datasets on various Auto-Encoder variants show that our loss is able to outperform the MSE loss and the Vanilla SSIM loss. We also provide reasons why our model is able to succeed in cases where the standard SSIM loss fails.

The MSE has several advantages: it is convex, differentiable, and fast to compute. In the image domain, the MSE has been a great reconstruction loss metric for Auto-Encoders (AE) (Hinton and Salakhutdinov 2006). Even though the MSE has proved to be the most successful loss for AEs, its major feature, measuring error of each dimension independently, does not align with the assumption in the actual model as we point out in the case of convolutional Auto-Encoders (C-AE). A C-AE uses its sliding convolutional filters to extract features such as ears or eyes, which requires to sample an area rather than individual pixels to encode meaningful information. Thus C-AE assumes that pixels are depending on its surrounding pixels to express information which is not aligning with the assumption of MSE. Another problem with using the MSE as reconstruction loss is that it measures absolute pixel differences rather than the structural difference which is more aligned with human's perception.

Existing works that seek alternative reconstruction loss mainly utilize the SSIM proposed by (Wang et al. 2004). However, despite the fact that it has shown its strength in measuring the image similarity, it has some limitations including not as effective in RGB images as it does in grey-scale images, and not being sensitive to luminance. Some works seek to alleviate its limitations by combining the SSIM score with the Mean Absolute Error (MAE) or the MSE (Zhao et al. 2017), but those formulations lack straight-forward explanations.

In the following sections we show SSIM's limitations by providing mathematical intuitions and empirical results. We then propose LWSSIM, a revised SSIM loss, and provide intuitive justifications for our formulation and experiment results.

## The SSIM and Its Limitations

The SSIM is originally constructed as an image quality measure with respect to the human perception rather than absolute differences measured by metrics such as the MSE or Mean Absolute Error (MAE).

To express the SSIM formally, according to (Wang et al. 2004), consider a pair of images $\{x, y\}$ of sizes $m \times n$, we want to measure three aspects of similarities according to human perception: luminance($l(x, y)$), contrast($c(x, y)$), and structure($s(x, y)$). Those are quantified according to the summary of relative measures including mean, variance, and co-variance measured under sliding windows of size $\xi \times \xi$ with step size of 1 on both horizontal and vertical directions.

For each sliding window, we measure base quantity for images $x$ and $y$ respectively. Then each perception sub-function is computed as follows:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (1)$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (2)$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (3)$$

Here $\{C_1, C_2, C_3\}$ are constants less than 1 to balance potential zero division issue. Usually, $C_3 = \frac{1}{2}C_2$.

To enforce independence among those measures, the final SSIM is constructed as the product of those metrics with exponential constant weights $\{\alpha, \beta, \gamma\}$ as:

$$SSIM(x,y;\xi) = l(x,y)^\alpha \cdot c(x,y)^\beta \cdot s(x,y)^\gamma \quad (4)$$

Notice that the luminance function is different from contrast and structure measures in that it uses mean to measure relative difference in luminance as oppose to absolute luminance difference. Mathematically, mean measures can output same value despite the range of pixel value in our measure window. In other word, the SSIM is not able to discriminate between a picture with high light-dark variance

and a picture with more consistent light and dark if the two pictures have same mean luminance value. Further, since the SSIM formulation uses multiplication between different sub-functions thus encourages independence among those factors, luminance's limitation is hard to be compensated by contrast or structure measures.

This is not a problem when measuring similarities between two images because the values of the two images are fixed. However, this is problematic when we want to use the SSIM as a loss function to guide the image reconstruction. Models with the SSIM as loss would eventually not able to learn to discard images generated with low luminance variance if other metrics are optimal.

## Proposed Structural Similarity Loss

We aim to adjust the existing SSIM metric to increase information acquisition in two folds. First, we replace the multiplication between luminance and the other with addition, and thus is able to remove independence between the three sub-functions to allow variance and co-variance to compensate mean. Secondly, we calculate the score under different filter size and calculate the weighted average among different level of filter sizes. This aims to alleviate the problem that the luminance metric is over "averaged" due to its large filter area. Different sized filters also allow the loss to capture different levels of localized information.

Formally, consider a pair of images $\{x, y\}$ of size $m \times n$ where $x$ is sampled from true distribution $p(x)$ and $y$ is sampled from the Auto-Encoder network using a corresponding $x$: $y = h_{\theta\phi}(x)$ where $\theta, \phi$ denotes parameters of encoder and decoder network respectively. For the $i^{th}$ level of similarity measure, we apply sliding window of different size $\xi_i \times \xi_i$ that moves pixel-by-pixel over the entire image. Then for each $\xi_i$ we have:

$$LWSSIM(x, y; \xi) = l(x, y)^\alpha + c(x, y)^\beta \cdot s(x, y)^\gamma \quad (5)$$

And for $I$ levels of different filter sizes we calculate the weighted sum with hyperparameters $\lambda_i$

$$LWSSIM(\mathbf{x}, \mathbf{y}) = \frac{1}{I} \sum_i \lambda_i LWSSIM(\mathbf{x}, \mathbf{y}; \xi_i) \quad (6)$$

## Experiments & Results

We applied our loss to replace the MSE part of loss functions in trending AE models including vanilla Auto-Encoder, Variational Auto-Encoder (VAE) (Kingma and Welling 2013), and MMD AE (Tolstikhin et al. 2017). All of the aforementioned models were implemented with convolutional layers in encoders and decoders. Observing figure 1, compared to images generated from models trained with MSE loss, we can see that the original SSIM loss enables the trained AE to generate less blurry images but with less vibrant colors resulted from discussed luminance issue across three channels. Our LWSSIM loss allows AE to generate images with higher levels of detail while preserving the color vibrancy. From quantative perspective, we see that the LWSSIM loss allows the model to generate images not only with higher structural similarity scores but also with lower MSE

Table 1: Reconstruction Quality Results on Celeb-A 64

| MODEL | SSIM | MSE |
|---|---|---|
| AE (MSE) | 0.80 | 0.0042 |
| AE (SSIM) | 0.86 | 0.0057 |
| **AE (LWSSIM)** | **0.89** | **0.0054** |
| VAE (MSE) | 0.53 | 0.0235 |
| VAE (SSIM) | 0.60 | 0.0244 |
| **VAE (LWSSIM)** | **0.59** | **0.0206** |
| MMD AE (MSE) | 0.81 | 0.0034 |
| MMD AE (SSIM) | 0.86 | 0.0062 |
| **MMD AE (LWSSIM)** | **0.88** | **0.0033** |



Figure 1: MMD AE reconstruction, MSE(left), SSIM(mid), LWSSIM(right). Notice the color loss in SSIM pics.

values most of the times, indicating that the LWSSIM is a good candidate to ensure both structural and absolute similarity.

## References

Hinton, G. E., and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *science* 313(5786):504–507.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Tolstikhin, I.; Bousquet, O.; Gelly, S.; and Schoelkopf, B. 2017. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13(4):600–612.

Zhao, H.; Gallo, O.; Frosio, I.; and Kautz, J. 2017. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging* 3(1):47–57.