ANALYSIS AND IMPROVEMENT OF IMAGE QUALITY EVALUATION METHODS IN GENERATIVE MODELS

Yaroslav Revera, Olha Peliushkevych

Ivan Franko National University of Lviv,
Department of Applied Mathematics and Informatics

yaroslav.revera@lnu.edu.ua

Introduction

Generative neural networks represent one of the most dynamic and rapidly developing areas in artificial intelligence. Beginning with the introduction of Generative Adversarial Networks (GAN) in 2014, followed by Variational Autoencoders (VAE) in 2019 and diffusion models in 2020, the field has evolved to innovative architectures like DALL-E and Midjourney, capable of creating realistic images from text descriptions and/or image inputs.

As these models evolved, so did image quality evaluation approaches. Simple metrics like mean square error proved inadequate since generative models often produce structurally different but still valid results. This led researchers to develop specialized metrics such as Inception Score (IS) and Fréchet Inception Distance (FID), which assess not only visual quality but also diversity and realism.

The importance of improving image quality metrics arises from the fast expansion of generative models across fields – from design to medical diagnostics to aerospace research. Current metrics often fail to keep up with generative models and miss some aspects, including alignment with human perception, semantic and structural consistency, diversity, and contextual relevance. Furthermore, even more problems arise if the generated data is domain-specific or unique.

Generative models and datasets

Despite the wide range of architectures, we focus on U-Net [1] and GAN [2]. U-Net was selected for this study due to its simplicity, popularity, stability, and relative accuracy across numerous applications.

The GAN architecture was proposed in 2014 and has significantly evolved with multiple updates in architecture, loss function formulation, and capabilities, as a result of contributions from numerous researchers.

GANs employ two neural networks, a Generator and a Discriminator, that learn through competition. The Generator produces images, while the Discriminator tries to identify them as real or fake.

In this work, we investigate GAN's variants CycleGAN [3] and pix2pix [4], which are widely recognized as high-quality and reliable generative models.

CycleGAN is a variation of GAN, which allows for unpaired image-to-image translation, incorporating cycle consistency loss to ensure that translating an image to another domain and back produces the original image.

Pix2pix, in turn, is a paired image-to-image translation approach that learns mapping between input and output image domains using U-Net as generator, combined with a PatchGAN discriminator to evaluate image authenticity through overlapping patches rather than entire images for adversarial loss.

We consider three paired image datasets: labels to facades [5], maps to aerial photos [4] and wind tunnel geometry to corresponding flow fields. The first two datasets require generative models to create realistic images that align with human perception. In contrast, the third dataset is domain-specific, where agreement with physical laws is rather important. The motivation for choosing this dataset stems from its industrial relevance, especially in fields like automotive engineering, where fast and accurate flow predictions for aerodynamic parts can drastically reduce time and costs.

Existing and proposed metrics

Evaluation of generative models often uses metrics like Inception Score (IS) [6], Fréchet Inception Distance (FID) [7] and Structural Similarity Index (SSIM), which are widely regarded as de facto standard for image evaluation. Metrics, such as IS and FID, heavily rely on the commonly accepted perception and expectations of images as they utilize the Inception v3 network trained on existing datasets of photos and illustrations to compare the feature distributions of generated images with those of images from the dataset. This concept makes them less reliable for domain-specific or unique images, where such features may be meaningless. Metrics like SSIM, while relying on measuring perceptual similarity (for example, luminance, contrast, and structure), struggle to capture deeper structural or semantic differences, especially when slight pixel shifts occur. Therefore, the problem with a lack of robust quality evaluation techniques or, in contrast, domain-specific metrics, remains acute.

In this work, we propose a generalized evaluation metric for images, which not only corresponds to visual perception but also integrates physical consistency for flow fields, as validated with dataset 3. We call such a metric enhanced Vigorous Features Fréchet Distance (eVFFD). Although it does not directly evaluate physical governing equations, eVFFD captures physical consistency through comparison of

enriched vector features extracted from images of the real and generated velocity fields from all the patches of images. The core of the metric is the Fréchet distance (see (1)) between the multivariate distributions of these features.

$$D^2 = \left\|\mu_r - \mu_g\right\|^2 + Tr\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{\frac{1}{2}}\right) \quad (1)$$

where μ and Σ refer to mean and standard deviation; subscripts r and g refer to real and generated data.

In this work, we propose the following components for the feature vectors:

- Local statistical characteristics: mean, standard deviation, skewness, kurtosis, and entropy.
- Gradient-based features: distribution of gradient magnitude, approximately reflecting spatial changes in the flow field.
- Texture features: extracted via Gabor filters, which detect oriented structures, capturing flow texture. This provides an additional level of sensitivity to the physical and geometric features of the flow structure.

Due to these components, the eVFFD metric captures local spatial structures, not only global statistics, indirectly reflects flow structure, gradients, and boundary layers, and works with scalar fields of velocity magnitude, without requiring the full velocity vector field, and the need to solve governing equations.

The improvement of eVFFD lies in two main aspects:

1. The specific combination of physically relevant features we selected.
2. The patch-based approach to feature analysis of local regions rather than the entire field at once (although patch-based image analysis is not new). We will show that such an approach is critical for accurate evaluation.

The output of eVFFD is a non-negative scalar indicating how much the generated field deviates from the real one. A value of zero indicates identical distributions, while higher values indicate worse matches in structure, statistics, or texture. Practical thresholds depend on the task and natural data variability.

Although developed for airflow fields, eVFFD also performs well on other datasets, as its selected features capture fundamental structural and textural characteristics applicable in broader contexts.

Results

We consider eVFFD with patch sizes of 16x16, 32x32, and 64x64. Smaller patch sizes are computationally expensive and do not correspond to the scale of images in our datasets, while a larger patch size misses important local structures.

In the map-to-satellite case, all metrics showed similar results, choosing CycleGAN, which is close to human perception. In labels-to-facades, both FID and eVFFD 64x64 evaluate the best generative model according to human perception. SSIM picked CycleGAN, even though it has a lot of artifacts.

For example, for the wind tunnel dataset, we demonstrate a more thorough case study below. In Image 1, we show the metrics' evaluation of images generated by different models, and in Image 2, we demonstrate an example from the dataset.

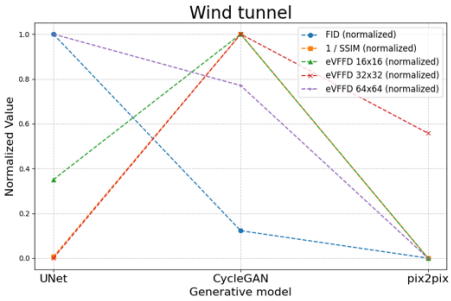

Image 1. Values of metrics for different generative models for the wind tunnel dataset

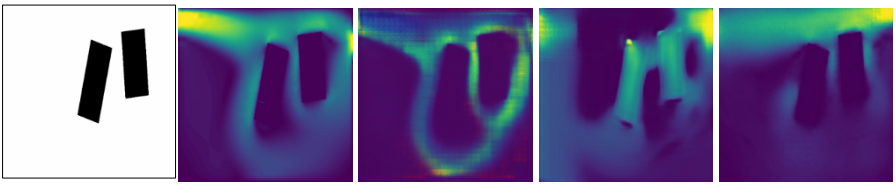

Image 2. Left to right: geometry (input), real field (simulations), generated images (output) by U-Net, CycleGAN, and pix2pix

As we can see, CycleGAN has the worst performance in terms of physical consistency, because it has positive velocity values at the areas corresponding to obstacles (black objects on the geometry image). However, pix2pix generated a realistic flow field, accurately reproducing major flow characteristics. eVFFD identified it correctly, but FID ranked CycleGAN as "close to best", which is a major mistake. It is important to note that the patch size in eVFFD is a crucial parameter and depends on the images and the size of features we want to investigate.

Conclusions

In the quickly advancing field of Generative AI, image evaluation metrics lag behind models, especially when it comes to unique domain-specific images. After investigating U-Net, CycleGAN, and pix2pix generative models on labels to facades, maps to aerial photos, and wind tunnel geometry to corresponding flow fields

datasets, we show that existing metrics, such as FID and SSIM, although widely used and considered a de facto standard, do not always reliably assess the visual and physical consistency of generated images. Therefore, we have proposed our enhanced metric, eVFFD, which is based on Fréchet distance, but instead of using abstract features, we focus on measuring the difference between the multivariate distributions of physically meaningful features, in particular local statistical characteristics, gradient and texture features. Such an approach allows for the correct evaluation of image quality in specific domains, where not only human perception but also underlying physical consistency are important.

In future work, we would like to explore the application of eVFFD to other domains, further integrating it into the training objective to improve the quality and physical consistency of generators.

References

1. O. Ronneberger. U-NET: Convolutional Networks for Biomedical Image Segmentation / O. Ronneberger, P. Fischer, & T. Brox. // Medical Image Computing and Computer-Assisted Intervention.– 2015.– pp 234-241.
2. Goodfellow I. Generative adversarial networks. / J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, & Y. Bengio // NIPS.– 2014.
3. J. Zhu. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks / J. Zhu, T. Park, P. Isola, A. Efros. // IEEE International Conference on Computer Vision (ICCV).– 2017.– pp. 2242-2251.
4. P. Isola. Image-to-Image Translation with Conditional Adversarial Networks. / P. Isola, J. Zhu, T. Zhou, & A. Efros. // IEEE Conference on Computer Vision and Pattern Recognition (CVPR).– 2018.– pp. 5967-5976.
5. R. Tyleček, Spatial Pattern Templates for Recognition of Objects with Regular Structure / R. Tyleček, & R. Šára // Lecture Notes in Computer Science.– 2013.– vol. 8142.
6. T. Salimans, Improved techniques for training GANs / T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, & X. Chen // NIPS.– 2016.– pp. 2234-2242.
7. M. Heusel. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium / M. Heusel., H. Ramsauer, T. Unterthiner, B. Nessler, & S. Hochreiter // NIPS.– 2018.– pp. 6629-6640.