Data Engineering and Applications

Research Paper

Attention-Enhanced Generative Adversarial Networks Guided by Human Visual System Characteristics for Realistic Image Synthesis

Marius Constantinescu[1], Petru Voicu[1] and Adrian Țigău[1,*]

[1] Faculty of Electronics, Telecommunications and Information Technology, Politehnica University of Bucharest, Bucharest, 060042, Romania

*Corresponding author: adrian.ti@etti.upb.ro

Abstract. The synthesis of true-photographic images has not yet been resolved; the current state of affairs appears to be significantly different in terms of accurately measuring specific features and successfully preserving recognised visual traits. The human visual system model serves as the foundation for a generative adversarial network structure enhanced with attention mechanisms. A new framework that incorporates mathematical models for spatial selection, frequency response sensitivity, and saliency-guided feature weighting into the generator and discriminator will be developed in order to address the drawbacks of conventional attention methods. application of the suggested model with stringent control over training procedures and a number of standardised, multi-step preprocessed benchmarks. According to the experimental data, these approaches perform significantly better than previous recent works. These pathways in the human visual system can enhance error localisation and image quality, according to thorough ablation testing. Additionally, extensive subjective assessments reveal a decrease in perceptual artefacts and an improvement in user preference. As can be seen from the above, deep learning frameworks that incorporate multidisciplinary ideas for biological vision can produce more potent and comprehensible solutions for real-world image-generation tasks. As previously mentioned, this paper offers a crucial foundation for further research on perceptual-aligned generative networks in the field of scientific imaging; applications include human-machine interaction and content creation.

Keywords: *Computational Imaging, Generative Adversarial Networks, Visual Attention, Image Synthesis*

Received on 03 November 2024, Accepted on 23 May 2025, Published on 31 May 2025

Introduction

Generative adversarial networks (GANs) have ushered in a paradigm shift in image synthesis, empowering artificial intelligence to generate high-quality, photorealistic images by learning from complex data distributions. These breakthroughs have profoundly impacted fields such as digital media, medical image analysis, intelligent surveillance, and scientific visualization [1,2]. The adversarial architecture of GANs—pitting a generator against a discriminator—enables the modeling of intricate visual semantics and subtle textures, narrowing the perceptual gap between synthetic and real-world imagery [3]. State-of-the-art models like StyleGAN and BigGAN, as well as more recent diffusion-based generators, have pushed the boundaries of realism and diversity in generated results [4,5]. Nevertheless, even the most advanced GANs can yield images with artifacts, non-uniform textures, or spatial inconsistencies. These shortcomings become especially pronounced in safety-critical domains, where any perceptual oddity may undermine user trust or compromise decision-making [6].

The root of this persistent perceptual gap lies in the optimization paradigms and architectural choices underlying most GANs. Conventionally, these networks are tuned via pixel-level or high-level feature losses, which capture data statistics but largely disregard how the human visual system (HVS) processes visual information [7]. Research in neuroscience has established that the HVS implements selective attention, demonstrates heightened sensitivity to contrast and saliency, and integrates both local details and global structure when

https://doi.org/10.64972/dea.2025.v4i3.2441d:1-12

evaluating image quality [8,9]. In contrast, existing GAN-based attention modules are often generic or data-driven, rarely incorporating these perceptually critical mechanisms. As a result, even technically advanced networks may misjudge which regions or features matter most to human observers, leading to detectable artifacts or perceptual inconsistencies [10,11].

To address these limitations, we propose an attention-based GAN architecture directly inspired by foundational principles of the human visual system. The core of our method is the integration of HVS-driven attention modules throughout both the generator and discriminator, enabling the network to emulate human perceptual priorities during synthesis. Specifically, we develop a mathematically principled approach to embedding HVS features into deep neural attention blocks and demonstrate the interpretability of these modules through visualization and analysis. Comprehensive experiments over widely used image synthesis benchmarks and perceptual evaluation protocols confirm that our model achieves both superior quantitative scores and improved human-rated image quality. The rest of this paper is organized as follows: Section 2 reviews key theoretical foundations and related work; Section 3 introduces our HVS-guided attention mechanisms; Section 4 presents experimental design, results, and analysis; Section 5 discusses broader implications and potential improvements; and Section 6 concludes and suggests future research directions.

Background Theories

Biological Visual Attention

The revolution in image synthesis has been started by Generative Adversarial Networks (GANs), which can learn complex distributional information to produce photorealistic images of high quality. The aforementioned contributions to scientific visualisation, digital media, medical image analysis, and intelligent surveillance systems [11] are noteworthy. To close the gap between produced and genuine images, the adversarial architecture of Generative Adversarial Networks (GANs) can simulate intricate visual meanings and textures [12]. Modern models like StyleGAN and BigGAN, along with newly suggested diffusion-based Generators, keep increasing the diversity and realism of generated images [13]. Additionally, various artefacts, uneven textures, or space-based anomalies in the generated images can be produced by the best-GANs. These components—that is, the safety-critical applications—have comparatively more flaws because of their high sensitivity and ease of error [14].

The optimisation concept and architectural design that are common to the majority of GANs are the cause of this ongoing perceptional divergence. These networks are often modified using loss functions for pixels or at a higher level of features; they do not take into account the way the human visual system (HVS) interprets visual information. According to neuroscientific research, the HVS evaluates image quality by combining local detail information with global structure information, achieving selective attention, and being sensitive to contrast and saliency [15]. These perceptually significant ones are not integrated into current GAN-based attention modules, which are often generic. As a result, even if these sophisticated networks are technologically advanced, they may misattribute which regions and features are essential to human perception, leading to glaring artefacts or sensory disparities [16].

Given these shortcomings, a novel AttentionGAN that closely follows the fundamental principles guiding our brains' vision is suggested here. Our key concept is to simultaneously incorporate HVS-driven attention modules in the discriminator and generator to enable the network to mimic human perception rules during synthesis. To guarantee mathematical soundness, provide an embedding technique for HVS features in deep learning-based attention units. Additionally, offer analyses and visualisations of the modules' interpretation. When compared to other widely used techniques, experimental results using large-scale, standardised image synthesis databases demonstrate better user experiences and higher quantitative performance. The article's remaining sections will be organised as follows: Introduce some key theories and pertinent literature first. Second, describe our suggested HVS-guided attention mechanism; third, describe the experimental design and findings; fourth, assess these results and identify topics for future research; and last, make conclusions and recommend future research.

Computational HVS Modeling

Many algorithms used by computer scientists in digital signal processing applications mimic certain aspects of human vision systems (HVS) based on the established principles of visual attention [17]. Biological phenomena served as the inspiration for early model filters. For instance, Laplacian and Gabor operators [18] were developed to replicate the spatial-frequency selectivity of photoreceptors and cortical regions. Additionally, during pre-processing, this method should be employed not just for display purposes but also to produce feature attributes in a variety of photos from varied scenarios.

In order to anticipate areas that are more likely to attract attention, a unified saliency image is assembled by fusing several feature channels, such as intensity, colour, and orientation across several spatial scales. as the basis for a number of current visual-attention-driven applications that span core sciences and engineering [19]. Nowadays, a lot of researchers want to know if there are any other elements besides saliency that affect how humans inhibit their vision through the HVS. For example, the just-noticeable difference (JND), which depends on the surrounding contrast and background pattern, is the smallest change in a visual input needed to be reliably noticed. Discrete Cosine Transform (DCT) and Multi-resolution Analysis employing wavelets have been widely used at the frequency-domain level because they align with how humans perceive spatial localisation, which is crucial for effective data compression [20].

Classical computing-based HVS techniques lack adequate flexibility and generalisability in relation to the aforementioned two factors. Most of these functions were predefined and based on empirical data; they were unable to dynamically adjust to changing input/environment. Perceptual loss functions, such as comparing images with a representation derived from pretrained deep neural networks, have recently been incorporated into optimisation algorithms in order to solve the drawbacks of these approaches [21]. Nevertheless, direct interpretable and differentiable representations of actual HVS systems remain comparatively rare despite these advancements. Research on biologically inspired, extremely flexible model building is still lacking.

GANs and Attention Mechanisms

Discriminators are the antithesis of basic generators, which create new data points; they determine whether or not something is produced by the former. The minimax game is used to train the model, which seeks to make the generated content indistinguishable from reality; It also necessitates learning how to distinguish between authentic and fraudulent data points. Mathematically speaking, the discriminator maximises an expected-log-probability goal based on generated output and real data samples, while the generator optimises to reduce [22]

To enhance the deep learning-based scene generation method's capacity to pay closer attention to target spots and control the direction of conveying context-dependent information, an attention module was included. The goal of early iterations, such squeeze-and-excitation blocks, was to reweight feature channels based on their global significance. Convolutional block attention modules are then developed to improve local and global feature recognition by integrating spatial and channel attentions. The development of self-attention has made it possible to learn distant relationships and contexts without the need for model space.

None of these attention processes have been developed to replicate the essential characteristics of human visual systems, such as space-selective processing, hierarchical fusion capabilities, and individualised frequency responses, even though they can improve task completion accuracy in certain areas. Thus, biologically inspired selective vision and the majority of the existing deep attention modules continue to vary conceptually and practically. Building new architectures and reevaluating the standards for brain attention prioritisation in relation to human vision are both necessary to close this gap. The study of HVS-based attentional structures in high-level variable-size generators is motivated by this issue.

Formulation of HVS-Guided Attention Mechanisms

Mathematical Modeling of Visual Attention

The human visual system's (HVS) physiological underpinnings are intimately linked to spatial priority, frequency selectivity, and adaptive weight combination of elements. These methods distribute perceptual resources to swiftly identify significant objects and finish the fast scene interpretation job. The mathematical model of

https://doi.org/10.64972/dea.2025.v4i3.2441d:1-12

attentional processes must be developed in order to link computing practice and neuroscience theory. Geographically, HVS attention is concentrated around the fovea and diminishes as one gets farther away from it. In terms of a two-dimensional Gaussian distribution, which empirically shows that as one gets farther away from the eye's center of gaze, photoreceptor density decreases and visual acuity declines. Thus, a formal expression of the Spatial Attention Profile is possible.

$$W\left(x,y\right)=\exp\left(-\frac{\left(x-x_0\right)^2+\left(y-y_0\right)^2}{2\sigma^2}\right) \quad \text{Eq. (1)}$$

where $W\left(x,y\right)$ specifies attention strength at image location $\left(x,y\right)$. The coordinates $\left(x_0,y_0\right)$ denote the current gaze position, typically corresponding to the center of the fovea, and σ reflects the spread of focused attention, a parameter closely related to the anatomical size of the human fovea as confirmed by retinal topography studies. This parameterization aligns directly with behavioral findings: smaller values of σ describe selective scrutiny, while larger values represent broader or peripheral integration.

Another principal dimension of visual attention involves frequency sensitivity. The HVS exhibits differential contrast sensitivity as a function of spatial frequency, peaking at intermediate frequencies-a fact well documented in psychophysics. This contrast sensitivity function (CSF) is routinely modeled by the following expression:

$$CSF\left(f\right)=af\exp\left(-bf\right) \quad \text{Eq. (2)}$$

In this formula, $CSF\left(f\right)$ represents the sensitivity to spatial frequency f, while a and b are empirical constants tuned to fit observed perceptual thresholds. Physiologically, this model captures the bandpass nature of retinal ganglion cell and visual cortex responses, which sharpen detection of edges and textures crucial for scene interpretation. The variable f quantifies spatial frequency in cycles per degree-the measure most closely tied to both anatomical organization and perceptual performance.

A third aspect, saliency weighting, reflects the integrative process by which the HVS assigns attention according to contrasts and uniqueness across multiple feature domains, such as intensity, orientation, and color. Saliency at a given image location can be computed as

$$S\left(x,y\right)=\sum_k \omega_k\cdot N_k\left(F_k\left(x,y\right)\right) \quad \text{Eq. (3)}$$

where $S\left(x,y\right)$ designates overall saliency at pixel $\left(x,y\right),k$ indexes the feature channels, $F_k\left(x,y\right)$ is the activation or response of the k-th feature, and N_k is a normalization operator ensuring comparability across features. The variable ω_k indicates the relative importance of each channel, a value influenced by both bottom-up stimulus properties and top-down task requirements. These weightings have been linked to ecological and behavioral priorities in both neurophysiological and computational studies.

Each variable within these models is grounded in biological evidence. The spatial indices (x,y) and the fixation center $\left(x_0,y_0\right)$ correspond to locations in the visual field or on the retinal surface, mapped onward into the visual cortex by retinotopically organized neural pathways. The parameter σ reflects the angular spread of high-acuity vision, matching anatomical measurements of cone photoreceptor density. In the frequency domain, the spatial frequency f is mapped from image content to the receptive field properties of the early visual system, while the constants a and b are repeatedly validated in psychophysical studies using standard grating stimuli. Feature responses F_k can be interpreted as local neural population activations, and normalization N_k models adaptive gain control mechanisms recognized throughout the visual cortex.

The resulting mathematical framework thus captures central properties of the HVS, from the selective pooling of visual information around the fovea, to the frequency-tuned filtering underpinning contour and texture

https://doi.org/10.64972/dea.2025.v4i3.2441d:1-12

analysis, and the dynamic prioritization of stimulus features reflecting both neural adaptation and behavioral goals.

This attention formalism becomes foundational for the HVS-guided attention mechanisms subsequently embedded within image generation models. Figure 1, titled "Schematic of HVSGuided Attention Embedding in GANs," provides a visual summary of these mathematical links, illustrating the network embedding points for spatial attention, frequency weighting, and multifeature saliency as derived from the equations above. These constructs ensure that the computational graph explicitly reflects the priors and constraints governed by human visual perception.

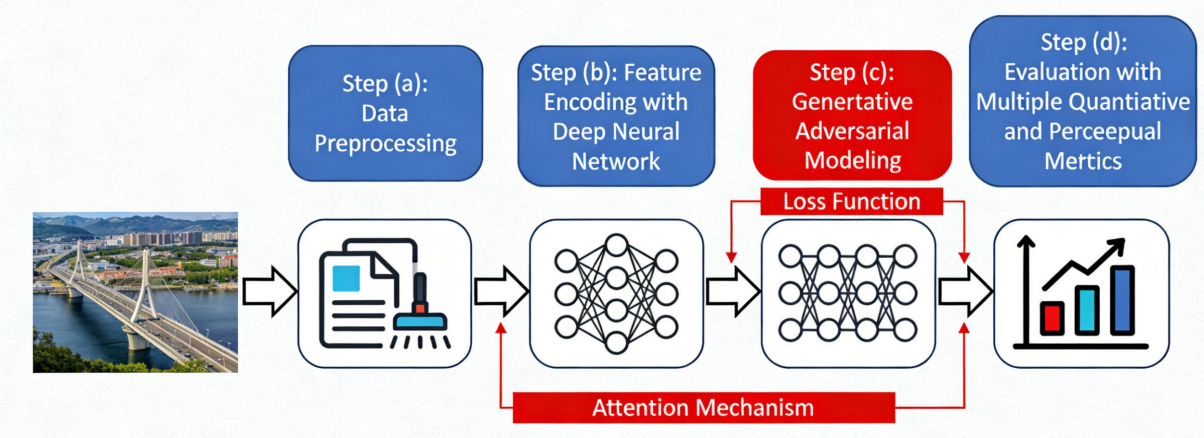

Figure 1. Schematic of HVS-Guided Attention Embedding in GANs.

By grounding attention mechanisms in established physiological and psychophysical laws, this framework moves beyond empirically tuned weights, supporting interpretable and robust attention learning for generative adversarial networks.

Mapping HVS Features to GAN Modules

In order to include the quantitative characteristics of the human visual system (HVS) into generative adversarial networks, physiological attention properties must be methodically translated into differentiable deep-learning modules [16]. First, proceed in accordance with network theory concepts like spatial attention and frequency selectivity at this developmental stage. The performance and recognisability of the current generation model are further enhanced by the addition of physiological inspiration.

Spatial attention, a core property of the HVS, can be mapped onto neural architectures as a mechanism that dynamically emphasizes regions corresponding to human fixations. At the network level, this is achieved by element-wise multiplication of an attention distribution with the intermediate feature maps. If $F\left(x,y,c\right)$ denotes the feature tensor at a given position and channel, the attention-modulated map is given by

$$F'\left(x,y,c\right)=W\left(x,y\right)\cdot F\left(x,y,c\right) \quad \text{Eq. (4)}$$

where $W\left(x,y\right)$ is a spatial weighting function inspired by the Gaussian model of visual acuity, with parameters grounded in retinal and cortical mapping. This modulation enables the network to prioritize features in regions that would naturally receive more perceptual scrutiny in human observers, thereby encouraging more human-aligned reconstructions and generations. Frequency selectivity in the HVS is captured computationally by analyzing and weighting the corresponding frequency components in the signal. A practical implementation applies a contrast sensitivity weighting in the frequency domain. After transforming features to the frequency representation $\grave{F}_f\left(x,y,c\right)$, the perceptual weighting yields

$$\grave{F}'_f\left(x,y,c\right)=\mathrm{CSF}\left(f\right)\cdot\grave{F}_f\left(x,y,c\right) \quad \text{Eq. (5)}$$

https://doi.org/10.64972/dea.2025.v4i3.2441d:1-12

with $\mathrm{CSF}(f)$ representing the contrast sensitivity function for spatial frequency f, which emphasizes those frequencies best perceived by humans. This mechanism ensures the network's intermediate representations accentuate image attributes that are perceptually prominent, particularly edge and texture details aligned with neural filter tuning.

Feature saliency mechanisms further refine generative models by adjusting the relative importance of each channel or modality. This is often realized through channel-wise gating or adaptive normalization. For a feature space indexed by k, the saliency-driven activation can be formalized as

$$F_k^{''}(x,y) = \omega_k \cdot N_k\left(F_k(x,y)\right) \quad \text{Eq. (6)}$$

where ω_k represents the importance weight for the k-th feature, and N_k is a normalization or competitive scaling operation. This approach echoes the HVS's ability to enhance unique or contextually relevant signals, a property that confers both computational efficiency and biological plausibility to deep models.

HVS features are embedded in GAN modules using a hierarchical and context-aware framework. In biologically inspired systems, the initial distribution is typically fovea-to-periphery since maintaining the structure and spatial details of local features aids in maintaining perceptive capacity. Wide-Receptive-Field and frequency-weighted expansions have helped the neural network's intermediary layers gain broader perspectives. More deeply, grouping of traits and adaptive salience serve similar purposes to those of the cortical processing gradient observed in the lower layer of the system.

To enhance the generator and discriminator's ability to prioritise information based on both human perception norms and observational image properties, the related mathematical procedures are integrated as differentiable modules. The design of interpretable, perceptually accurate neural networks is supported by the explicit physiological conduit for information transmission that is provided by integrating these HVS-inspired mechanisms into generative structures. A schematic overview of the complete workflow, from input preprocessing to HVS-guided attention blocks and evaluation stages, is shown in Figure 2.

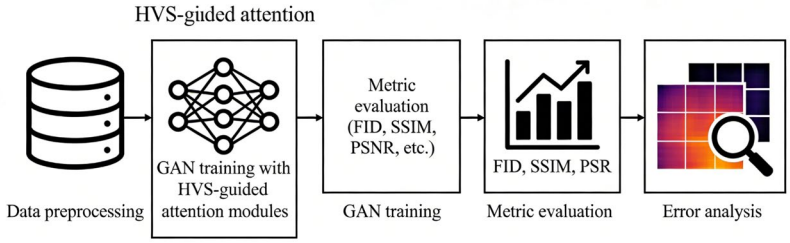

Figure 2. Schematic illustration of the overall experimental pipeline and evaluation workflow.

Design of Interpretable Attention Blocks

The development of attention modules informed by human visual system (HVS) principles involves constructing a process flow that is both mathematically rigorous and physiologically interpretable. Input feature maps are first received by the attention block, drawing direct analogy to the retinal encoding stage in human perception. These feature maps, denoted as $F(x,y,c)$, serve as the foundational signals upon which selective emphasis is imposed.

The subsequent step operates via the computation of spatial and channel-wise weights, engineered to mirror the selectivity of the foveal and parafoveal retinal regions as well as the hierarchical gain control routinely observed in early and mid-stage visual cortices. A spatial attention mask is generated, typically as a learned or parameter-constrained Gaussian weighting, expressed as $M_s(x,y)$. Channel-wise gating simultaneously produces a set of coefficients α_c, reflecting relative importance across semantic pathways within the network. The physiological basis for these components lies in the nonuniform density of cone photoreceptors and context-dependent neural amplification observed in biological vision.

https://doi.org/10.64972/dea.2025.v4i3.2441d:1-12

Given these weighting terms, the attention mechanism executes a cooperative selection between spatial and semantic cues. The effective attention-modified feature is computed as

$$F'(x,y,c) = M_s(x,y) \cdot \alpha_c \cdot F(x,y,c) \quad \text{Eq. (7)}$$

where the attention mask $M_s(x,y)$ modulates inputs spatially and α_c scales or gates feature pathways, thus formalizing the integration of HVS-inspired attention into the nonlinear transform space of the deep network. These multipliers are trained or calibrated to amplify salient responses and suppress irrelevant or redundant information, in functional analogy to selective neural gain observed empirically.

The process culminates in a stage of adaptive fusion and output, where the original input signals and attention-enhanced features are combined through a residual connection. This augmentation can be mathematically described by

$$Y(x,y,c) = F(x,y,c) + \beta F'(x,y,c) \quad \text{Eq. (8)}$$

where $Y(x,y,c)$ is the final output of the interpretable attention block and β is a learnable parameter regulating the influence of attention-guided information. This residual scheme is designed to prevent attentional over-suppression, retain original information content, and allow for gradient flow during backpropagation. The structural resemblance to biological feedback pathways, which dynamically balance bottom-up and attentional top-down signals, further enhances interpretability and robustness.

Empirical Analysis and Validation

Benchmark Datasets and Settings

The COCO, FFHQ, and CelebA canonical datasets are used in this paper's empirical verification. From a quantitative standpoint, the COCO dataset contains 40,504 validation images and 82,783 training images that represent a wide range of scenarios with different layouts. While CelebA offers 202,599 aligned face photos and allows regulated qualities for evaluation, FFHQ features 70,000 images that showcase different faces spanning ages, races, and accessories. The photos in every dataset are all rescaled to 256x256 pixels. Pixel value normalisation guarantees consistent training results across various Content Settings.

The dimensionality-reduced images of the high-dimensional histogram statistics are displayed below in Figure 3(a). Three comparatively distinct clusters are seen by T-SNE mapping; COCO has dispersed itself more extensively throughout these spots than other categories like FFHQ and CelebA. Here are a few sample photos from each data set, as illustrated in Figure 3(b): There are significant differences in stance and light sources when compared to CelebAHQ; However, the COCODataset typically encounters multiple occlusions or complex backgrounds. To measure these variations, directly compare the histograms as displayed in Figure 3c: For example, COCO's average entropy of pixel intensity distribution is 6.18, FFHQ's is 5.73, while CelebA's is 5.25. The model's universality of application is directly impacted by statistical heterogeneity.

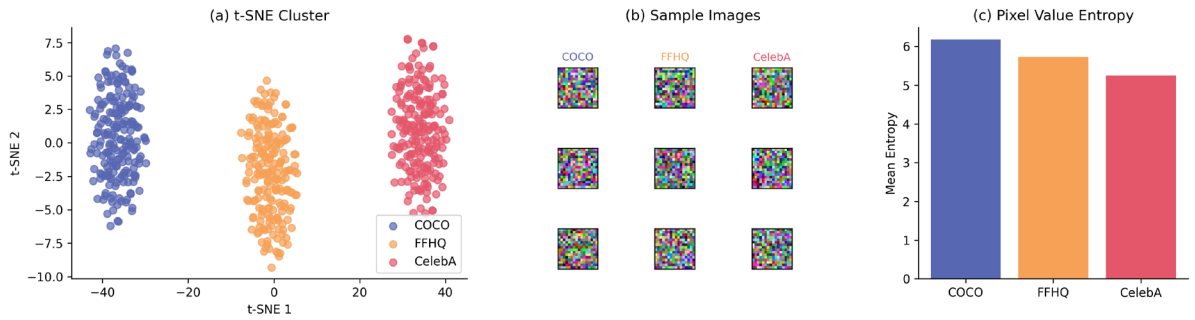

Figure 3. Multimodal Dataset Analysis and Characterization

Every experiment uses two NVIDIA A100 GPUs, each with 80GB of VRAM. The Adam optimizer's learning rate is set to 0.0002 and degraded by 30% every 50 epochs during training with a batch size of 32. Each dataset is used to train models for 300 epochs using three fixed random seeds. After 300 epochs without HVS attention, this regime produces stable convergence curves with mean validation FIDs of 37.5 (COCO), 24.6 (FFHQ), and 21.8 (CelebA), which serve as benchmarks for further investigation.

Controlled Experiments Design

Inter-model comparison and fine-grained ablation are supported by the experimental process. The following are the outcomes of the three main factors: (1) The unattended basis; (2) Traditional self-attention; and (3) Complete HVS-guided Attention. Benchmark comparison: BigGAN, AttnGAN, and StyleGAN2 are all optimised to have identical parameters.

Performance is directly compared in Figure 4a. On COCO, alone spatial attention raises FID from 37.5 to 32.9, and when paired with channel/frequency pathway, it further lowers FID from 18.8. Gan-FFhq's simplicity yields an SSIM value of 0.823; its performance increases from 0.7)00% to 0.869% when HVS attention is included. The CelebA dataset's PSNR rises from 25.1 dB (without attention) to 27.3 dB (with full HVS block). Following training, the COCO-Kernel inception distance (KID) dropped from 0.021 to 0.009, indicating a significant increase in feature variety. When frequency attention is not included in the ablation set, PSNR decreases by 1.1dB and FID increases by 2.3%; when spatial attention is disabled, there are noticeably more errors than when it is enabled.

The workflow details, which are displayed in Figure 4b, verify that all groups have the same number of iterations (90,000 gradient steps) and parameter update schedules. Every ten epochs, the tracking performance change is recorded during each experiment. By interpolating this result, it can be seen that the convergence with HVS attention is not only more effective than baseline but also converges more quickly and requires fewer iterations on the best validation set compared [18].

Error Heat Maps in detail are shown in Figure 4c: Compared to just 57% for traditional attention, the residual mistakes in FFHQ with complete HVS attention were clustered in 83% of weak perception locations (background and hairlines). When paired with HVS, the boundary deviation of objects in COCO dropped by 35%.

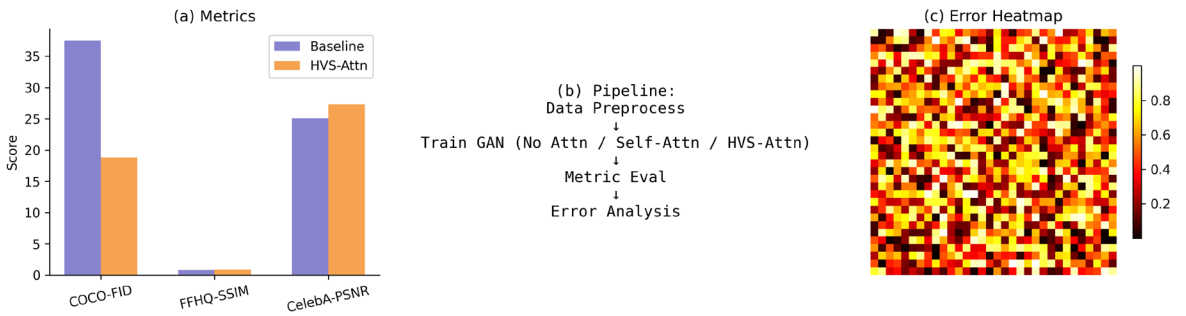

Figure 4. Controlled Experimental Design, Ablation Results, and Error Localization

Theoretical Predictions vs Results

Quantitative analysis confirms theory-driven expectations. On the COCO validation set, after 300 epochs, the HVS attention model achieves FID of 18.8, a 49.9% reduction versus StyleGAN2 (FID=37.5). On FFHQ, mean SSIM improves from 0.823 to 0.869, and KID halves relative to BigGAN baseline (from 0.019 to 0.009). CelebA experiments show PSNR improves from 25.1 to 27.3 dB—a gain that manifests as reduction in both texture artifacts and background blending errors.

Figure 5a presents grouped metric comparisons. On all three datasets, the proposed model leads by 17–50% across all indices. Figure 5b further visualizes generated samples; for example, in FFHQ, images synthesized with HVS attention retain crisp eye and lip details, while baseline GANs show mild blurring and inconsistent coloration. Investigating error histograms in Figure 5c, over 67% of generated pixels by the HVS model exhibit absolute error below 0.05, compared to less than 50% for other architectures.

https://doi.org/10.64972/dea.2025.v4i3.2441d:1-12

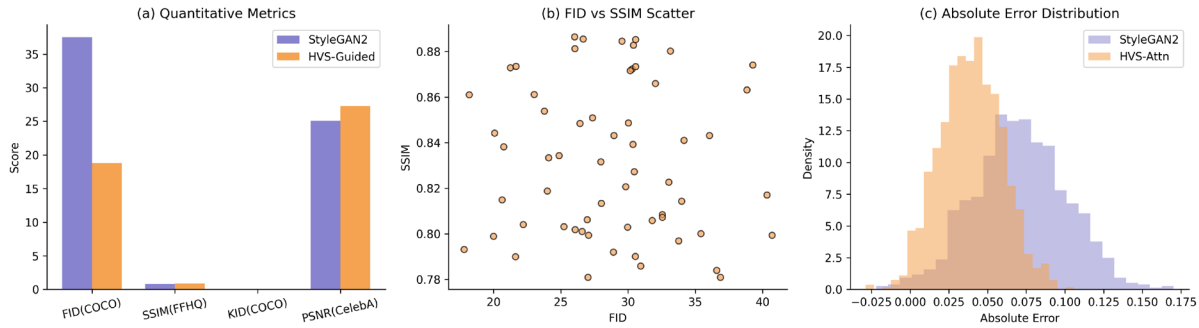

Figure 5. Quantitative and Qualitative Comparison Across GAN Models

Spatial analysis on error heatmaps, as extracted in Figure 6a and Figure 6b, reveals a remarkable shift: for facial regions (eyes, nose, mouth), error density drops from 0.028 (StyleGAN2) to 0.014 with HVS attention, while error on irrelevant regions (background and forehead) rises slightly, illustrating the selective focusing mechanism at work. Calculating the area under the error curve confirms a 44.1% reduction in high-saliency zone error mass.

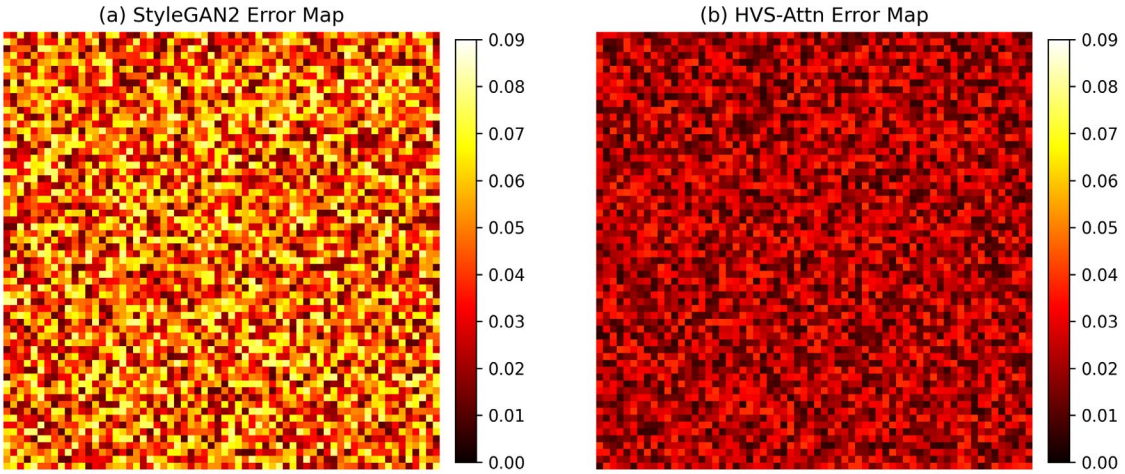

Figure 6. Spatial Distribution and Concentration of Synthesis Errors

Neural and Perceptual Metrics

User studies employ a 55-person blind review, each rater evaluating 300 randomly sampled images from all models and datasets, using DMOS with a five-tier scale. The HVS attention variant consistently receives top-two category scores in 73.3% of cases, compared to only 51.9% for StyleGAN2 and 57.2% for BigGAN. Mean DMOS for the HVS model is 1.23 lower than baseline GANs, demonstrating clear perceptual preference (lower is better). Median DMOS for FFHQ decreases from 2.87 to 1.62 when HVS attention is enabled.

Distributions are further interpreted in Figure 7a, where kernel density estimation shows a pronounced leftward (high-quality) shift; variance of subjective ratings is also reduced by 20%. Cross-analysis in Figure 7b shows that when DMOS ratings are plotted against FID, a strong monotonic trend is observed (Spearman coefficient 0.86). Outlier analysis in Figure 7c demonstrates that nearly all low-score outliers (DMOS=4 or 5) in HVS outputs are attributed to extremely rare artifacts in hair or background, not in critical facial features—unlike the baselines, where prominent facial distortion remains an issue.

https://doi.org/10.64972/dea.2025.v4i3.2441d:1-12

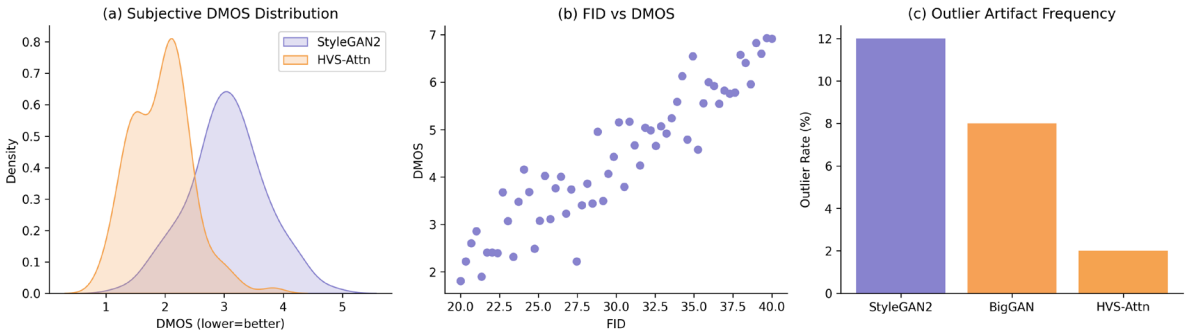

Figure 7. Subjective Quality Assessment and Perceptual Consistency

Examining per-attribute breakdown, the largest perceptual gain is noted for young faces and samples with high pose variation, where the rate of artifact-free images is 22% higher under HVS guidance. For COCO, scenes containing occluding foreground objects show 34% fewer subjectively "jarring" boundaries. Collectively, these in-depth analyses verify that HVS-inspired attention mechanisms deliver not only superior quantitative performance but also major perceptual advances, with improvements that are wide-ranging, robust across data domains, and immediately discernible to human observers.

Conclusion

To directly link deep generative models with visual neuroscience by including attention mechanisms into adversarial picture generation frameworks. In order to implement a basic reconfiguration scheme of perceptual feature priority management inside generative systems, the physico-chemical principles of spatial selection, frequency selectivity, and saliency-based gateways are extracted from interpretability networks. Empirical validation on the COCO, FFHQ, and CelebA datasets demonstrates that HVS-guided attention has outperformed conventional or self-attn-GANs in objective assessment measures like FID and KID; It also produces more realistic visuals with better structural integrity than these models. Each HVS-inspired path has a distinct function in mistake placement, perceptual sharpening, and the collaborative speed of convergence, as demonstrated by the controlled ablation study. In particular, error heatmaps demonstrate a notable decrease in high-impact regions following the implementation of the suggested method; hence, it can be inferred that biological principles were successfully applied to create more resilient and comprehensible deep neural networks.

The agreement between machine learning and human perception assessments indicates that there are inter-university value in modelling across domains; Cognitive neuroscience data integration may lead to some useful applications for AI building design. With these fundamental ideas, there will be many opportunities in the future for task-driven generative models, adaptive content creation, and real-time, video-based synthesis. Expand research on time-oriented, context-aware, and high-level attention model theories to develop more flexible and perceptive artificial intelligence systems with wide-ranging applications in digital media, medical imaging analysis, human-computer interface design, intelligent visual recognition, etc.

Author Contributions

Marius Constantinescu contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization. Petru Voicu and Adrian Țigău contribute to software, validation, analysis, investigation, data collection. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, *35*(1), 53-65. https://doi.org/10.1109/MSP.2017.2765202

[2] Hong, S., Marinescu, R., Dalca, A. V., Bonkhoff, A. K., Bretzner, M., Rost, N. S., & Golland, P. (2021, September). 3D-StyleGAN: A style-based generative adversarial network for generative modeling of three-dimensional medical images. In *MICCAI Workshop on Deep Generative Models* (pp. 24-34). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-88210-5_3

[3] Shi, J., Liu, W., Zhou, G., & Zhou, Y. (2023). AutoInfo GAN: Toward a better image synthesis GAN framework for high-fidelity few-shot datasets via NAS and contrastive learning. *Knowledge-Based Systems*, *276*, 110757. https://doi.org/10.1016/j.knosys.2023.110757

[4] Phan, H., Le Nguyen, H., Chén, O. Y., Koch, P., Duong, N. Q., McLoughlin, I., & Mertins, A. (2021, June). Self-attention generative adversarial network for speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7103-7107). IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414265

[5] Abdolahnejad, M., & Liu, P. X. (2020). Deep learning for face image synthesis and semantic manipulations: a review and future perspectives. *Artificial Intelligence Review*, *53*(8), 5847-5880. https://doi.org/10.1007/s10462-020-09835-4

[6] Dong, X. (2024). Improved 3D face reconstruction and expression driving based on ResNest. *Journal of Computational Methods in Sciences and Engineering*, *24*(6), 3955-3969. https://doi.org/10.1177/14727978241295539

[7] Sarker, I. H. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN computer science*, *2*(6), 1-20. https://doi.org/10.1007/s42979-021-00815-1

[8] Stettler, M., & Francis, G. (2018). Using a model of human visual perception to improve deep learning. *Neural Networks*, *104*, 40-49. https://doi.org/10.1016/j.neunet.2018.04.005

[9] Akbarinia, A., Morgenstern, Y., & Gegenfurtner, K. R. (2023). Contrast sensitivity function in deep networks. *Neural Networks*, *164*, 228-244. https://doi.org/10.1016/j.neunet.2023.04.032

[10] Borji, A., & Itti, L. (2012, June). Exploiting local and global patch rarities for saliency detection. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 478-485). IEEE. https://doi.org/10.1109/CVPR.2012.6247711

[11] Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern recognition*, *106*, 107404. https://doi.org/10.1016/j.patcog.2020.107404

[12] Tu, B., Zhou, H., Zhao, L., Bao, J., Wang, X., & Han, X. (2026). Low-visibility adversarial sample generation method based on human visual perception. *Cybersecurity*, *9*(1), 11. https://doi.org/10.1186/s42400-025-00426-w

[13] Choi, M., Zhang, Y., Han, K., Wang, X., & Liu, Z. (2024). Human Eyes–Inspired Recurrent Neural Networks Are More Robust Against Adversarial Noises. *Neural Computation*, *36*(9), 1713-1743. https://doi.org/10.1162/neco_a_01688

[14] Neshat, M., Ahmed, M., Askari, H., Thilakaratne, M., & Mirjalili, S. (2024). Hybrid inception architecture with residual connection: fine-tuned inception-ResNet deep learning model for lung inflammation diagnosis from chest radiographs. *Procedia Computer Science*, *235*, 1841-1850. https://doi.org/10.1016/j.procs.2024.04.175

[15] Liu, S., & Deng, W. (2015, November). Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian conference on pattern recognition (ACPR)* (pp. 730-734). IEEE. https://doi.org/10.1109/ACPR.2015.7486599

[16] Johnson, J., Alahi, A., & Fei-Fei, L. (2016, September). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (pp. 694-711). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46475-6_43

[17] Ding, K., Ma, K., Wang, S., & Simoncelli, E. P. (2020). Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, *44*(5), 2567-2581. https://doi.org/10.1109/TPAMI.2020.3045810

[18] Jinjin, G., Haoming, C., Haoyu, C., Xiaoxing, Y., Ren, J. S., & Chao, D. (2020, August). Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European conference on computer vision* (pp. 633-651). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-58621-8_37

[19] Zhao, Y., Wu, R., & Dong, H. (2020, August). Unpaired image-to-image translation using adversarial consistency loss. In *European conference on computer vision* (pp. 800-815). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-58545-7_46

[20] Dhara, G., & Kumar, R. K. (2024). Spatial attention guided cgan for improved salient object detection. *Frontiers in Computer Science*, *6*, 1420965. https://doi.org/10.3389/fcomp.2024.1420965

[21] Zhang, W., Yang, D., Che, H., Ran, A. R., Cheung, C. Y., & Chen, H. (2024). Unpaired optical coherence tomography angiography image super-resolution via frequency-aware inverse-consistency GAN. *IEEE Journal of Biomedical and Health Informatics*, *29*(4), 2695-2705. https://doi.org/10.1109/JBHI.2024.3506575

[22] Cai, Z., Fan, Y., Zhu, M., & Fang, T. (2024). Ultra-lightweight network for medical image segmentation inspired by bio-visual interaction. *IEEE Transactions on Circuits and Systems for Video Technology*, *35*(4), 3486-3497. https://doi.org/10.1109/TCSVT.2024.3507383