arXiv:2502.07753v1 [cs.CV] 11 Feb 2025

Direct Ascent Synthesis: Revealing Hidden Generative Capabilities in Discriminative Models

Stanislav Fort [1] **Jonathan Whitaker** [2]

Abstract

We demonstrate that discriminative models inherently contain powerful generative capabilities, challenging the fundamental distinction between discriminative and generative architectures. Our method, Direct Ascent Synthesis (DAS), reveals these latent capabilities through multi-resolution optimization of CLIP model representations. While traditional inversion attempts produce adversarial patterns, DAS achieves high-quality image synthesis by decomposing optimization across multiple spatial scales (1×1 to 224×224), requiring no additional training. This approach not only enables diverse applications – from text-to-image generation to style transfer – but maintains natural image statistics ($1/f^2$ spectrum) and guides the generation away from non-robust adversarial patterns. Our results demonstrate that standard discriminative models encode substantially richer generative knowledge than previously recognized, providing new perspectives on model interpretability and the relationship between adversarial examples and natural image synthesis.

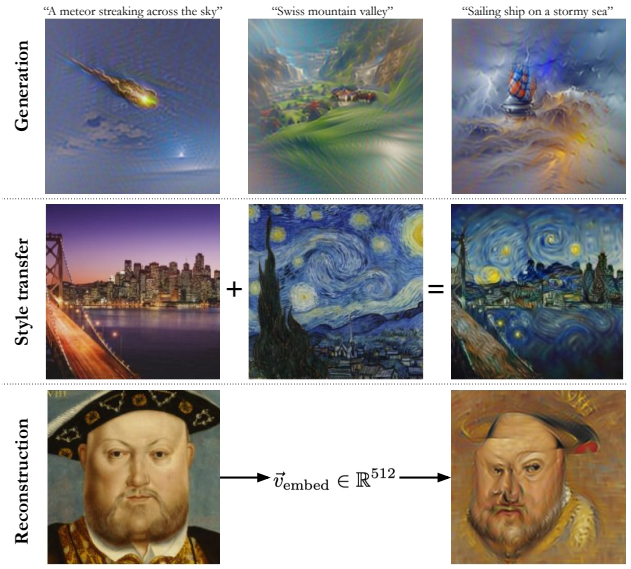

Figure 1. Direct Ascent Synthesis generates high-quality images by optimizing multi-resolution components to match CLIP embeddings, without any generative training. Unlike standard adversarial optimization that produces noise-like patterns, our approach reveals that pretrained discriminative models contain rich generative knowledge accessible through careful optimization. It can be used for a variety of image manipulations, such as style transfer and image reconstruction from a low-dimensional embedding.

1. Introduction

Machine learning has traditionally relied on a fundamental dichotomy: discriminative models map inputs to semantic representations, while generative models synthesize data from learned latent spaces. This separation has driven remarkable progress, from GANs (Goodfellow et al., 2014) to diffusion models (Ho et al., 2020; Rombach et al., 2022). However, these approaches require extensive training on large datasets, raising questions about whether such complex training procedures are fundamentally necessary for high-quality generation.

We challenge this dichotomy with Direct Ascent Synthesis (DAS), demonstrating that discriminative models implicitly encode rich generative knowledge accessible through careful optimization. The key challenge is that while discriminative models excel at mapping images to representations ($f : I \to v$), inverting this process ($f^{-1} : v \to I$) typically produces degenerate results. When optimizing an image I^* to match a target representation v, the result often achieves near perfect mathematical alignment ($f(I^*) \approx v$) while appearing as meaningless noise to human observers. This phenomenon, first noted in the context of adversarial examples (Goodfellow et al., 2015), reveals a fundamental tension between representation matching and perceptual quality. Trying to invert the latent representation back to a synthetic image was also the original motivation that led to the discovery of adversarial attacks.

Our key insight is that this degeneracy can be broken by

[1]Independent Researcher [2]Answer.AI. .

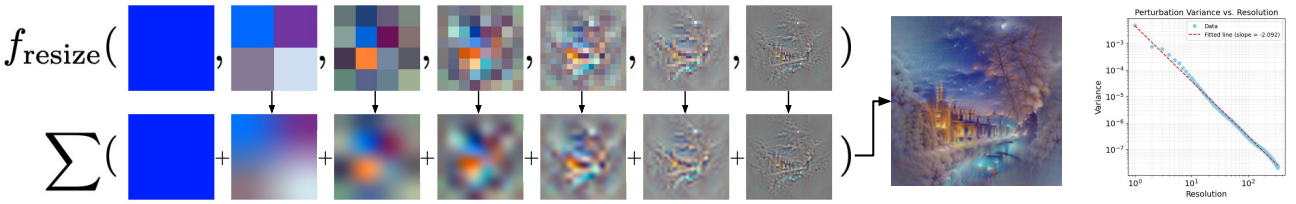

Figure 2. Multi-resolution decomposition enables training-free image synthesis. Left: An image is expressed as a sum of components at increasing resolutions, from 1×1 to 224×224. Middle: The components are optimized simultaneously to maximize CLIP embedding similarity with a target description, producing coherent images without generative training. Right: The power spectrum of generated images follows a $1/f^2$ distribution (slope ≈ -2), characteristic of natural images. This demonstrates that our multi-resolution prior effectively guides optimization toward perceptually valid solutions.

decomposing the optimization across multiple resolutions. This multi-scale decomposition provides natural regularization that aligns with human visual priors, preventing degenerate high-frequency solutions while enabling explicit control over the generation process.

This simple approach produces high-quality, semantically meaningful images without any training, suggesting that pretrained discriminative models contain richer generative capabilities than previously recognized. DAS requires only seconds of computation on a single GPU for inference (and no compute for generative training at all), challenging assumptions about the necessity of extensive generative training. Beyond practical benefits, our work provides insights into the fluid boundary between discrimination and generation in deep neural networks—perhaps both types of models learn similar underlying representations, just accessed in different ways.

Our approach reveals a surprising connection between adversarial examples and image synthesis. The same optimization process that typically produces adversarial patterns can be redirected toward meaningful generation through appropriate regularization. This suggests that adversarial examples may not be a fundamental limitation of discriminative models, but rather a symptom of optimization that ignores natural image structure (this is explored in (Fort & Lakshminarayanan, 2024)).

We validate DAS through experiments on image generation, controlled modification, reconstruction, style transfer, and inpainting tasks. Our results demonstrate that combining discriminative representations with appropriate optimization priors enables high-quality synthesis without the computational and data requirements of traditional generative training.

2. Related Work

2.1. The Evolution of Image Synthesis

Image synthesis has traditionally followed two parallel tracks. The generative track progressed from VAEs (Kingma & Welling, 2022) and GANs (Goodfellow et al., 2014) to diffusion models (Ho et al., 2020; Rombach et al., 2022), achieving remarkable quality through increasingly complex training. The discriminative track revealed rich internal representations through feature visualization (Yosinski et al., 2015; Nguyen et al., 2016; Olah et al., 2017) and adversarial examples (Goodfellow et al., 2015), while models like CLIP (Radford et al., 2021) demonstrated that discriminative training can capture general visual concepts.

2.2. Bridging Discrimination and Generation

Several works have hinted at deeper connections between these approaches. Feature inversion methods (Mahendran & Vedaldi, 2014) showed that discriminative representations contain generative information, though with limited quality. Analysis of GAN discriminators (Bau et al., 2019b) revealed latent spaces similar to generators, suggesting common representational principles. The success of optimization-based synthesis through techniques like deep image prior (Ulyanov et al., 2018) and neural style transfer (Gatys et al., 2016) demonstrated that careful optimization can sometimes replace explicit generative training.

More recently, the release of OpenAI's CLIP models (Radford et al., 2021) sparked a series of experiments in the open-source community that used CLIP similarity to a target text prompt to guide optimization in the latent space of various GAN generators. In particular, VQGAN-CLIP (Crowson et al., 2022) was used extensively for creative image generation and editing, and early practitioners quickly discovered the value of augmentations for improving and stabilizing the optimization process.

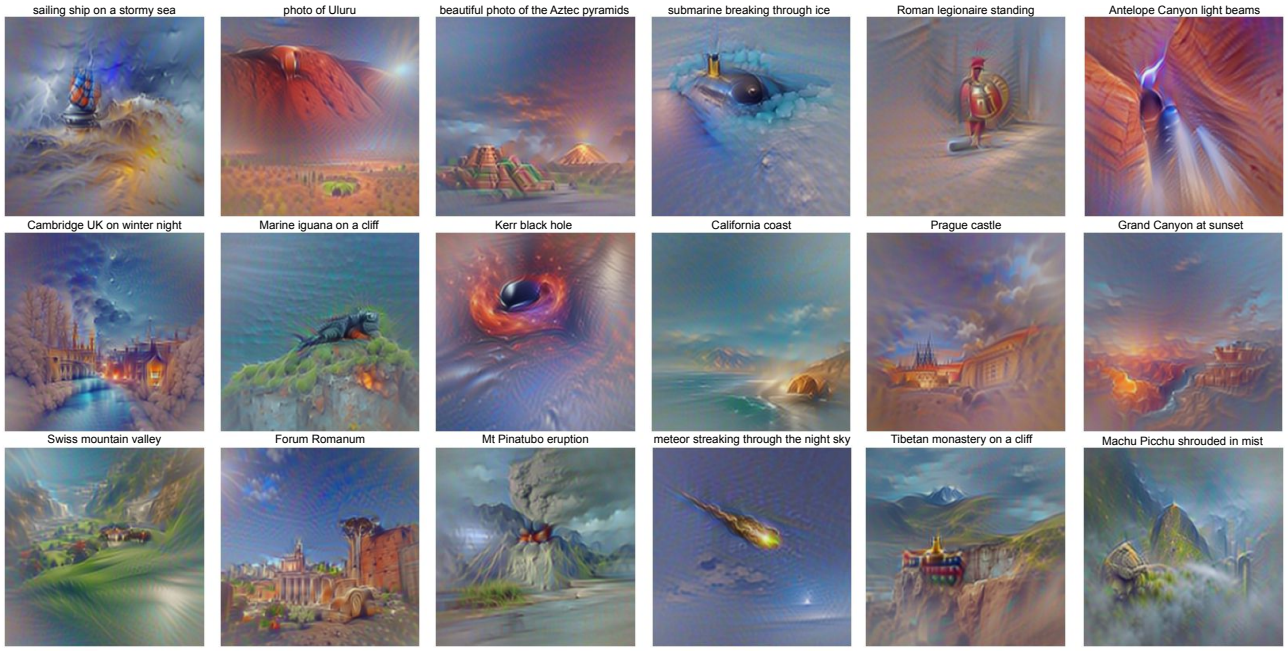

Figure 3. Diverse generations from Direct Ascent Synthesis across a range of concepts and styles. Results were obtained by optimizing against an ensemble of three CLIP models, with prompt augmentation to control image aesthetics: discouraging text generation (-0.3 × "Optical Character Recognition"), enhancing rendering quality (0.3 × "octane render, unreal engine, ray tracing, volumetric lighting"), and preventing image stacking (-0.3 × "multiple exposure").

2.3. Adversarial attacks on large models

Despite early hopes to the contrary (especially due to scaling, e.g. Dehghani et al. (2023)), large models still suffer from adversarial examples. Fort (2021a;b) shows that OpenAI CLIP models (Radford et al., 2021) can be fooled by small, easy-to-find, targeted, pixel-level modifications to the input image. Even very robust out-of-distribution detectors based on large scale pretrained models (Fort et al., 2021) suffer from an equivalent brittleness under targeted attacks (Fort, 2022). Transferable adversarial image attacks on proprietary models such as GPT-4, Claude and Gemini were first constructed in Fort & Lakshminarayanan (2024). While there have been dedicated approaches improving adversarial robustness on small datasets (e.g. Madry et al. (2019)), no solution has yet emerged at scale.

2.4. The Role of Multi-Scale Processing

The importance of multi-scale representations spans both classical and modern approaches (Lindeberg, 1994), from Gaussian pyramids (Burt & Adelson, 1983) to recent architectures with explicit multi-scale processing (Fort & Lakshminarayanan, 2024). This aligns with cognitive science findings that human visual processing operates across multiple spatial frequencies (Jeantet et al., 2018). Our work builds directly on these insights by showing that multi-resolution optimization can bridge the gap between discriminative and generative processes.

The most directly related work is Whitaker (2022), which independently explored similar ideas of optimization-based image synthesis in their open source project.

3. Multi-Resolution Optimization for Image Synthesis

The foundation of our approach lies in understanding how natural images are structured across scales. Since the development of Gaussian and Laplacian pyramids (Burt & Adelson, 1983), multi-resolution decomposition has been a powerful tool for analyzing images, revealing how information and statistics are organized across spatial frequencies (van der Schaaf & van Hateren, 1996). We extend these classical insights to guide generative optimization in deep neural networks.

Our key innovation is reformulating image synthesis as simultaneous optimization across multiple scales:

$$P(I) = \{I_r | r \in \rho\}, \text{ where } I_r = \text{rescale}_r(I) \quad (1)$$

This decomposition serves several crucial purposes: 1) Provides natural regularization by enforcing consistency across scales, 2) Captures semantic information at appropriate res-

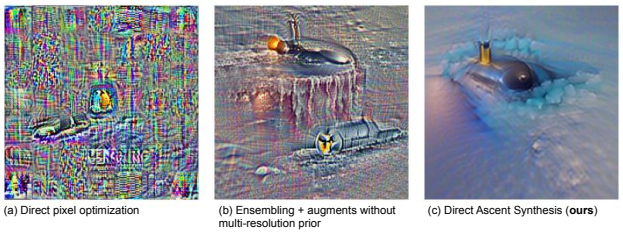

Figure 4. Ablation study demonstrating how different components of Direct Ascent Synthesis contribute to coherent image generation. Left: Direct pixel optimization yields adversarial patterns typical of model inversion attacks. Middle: Adding augmentations and model ensembling begins to impose structure but still lacks coherence. Right: Our complete approach with multi-resolution prior produces natural, interpretable images. This progression reveals how careful regularization can transform the degenerate solutions of model inversion into meaningful image synthesis.

olutions, 3) Prevents degenerate high-frequency solutions characteristic of adversarial examples.

While traditional adversarial optimization often produces noise-like patterns by exploiting single-scale processing, our approach encourages consistency across the natural scale hierarchy of visual information. This aligns with both human visual processing—where different neural populations respond to features at different scales—and recent findings that multi-resolution processing improves neural network robustness (Fort & Lakshminarayanan, 2024).

The optimization objective becomes:

$$I^* = \arg\min_{P_1,\ldots,P_R} \mathcal{L}\left(f\left(\sum_{r\in\rho} \text{resize}_{224}(P_r)\right), v\right) \quad (2)$$

where P_r represents image components at resolution $r \times r$, and $\mathcal{L}$ measures representation similarity. This formulation automatically encourages solutions that respect the statistical structure of natural images while maintaining semantic coherence across scales.

This connection between scale-space consistency and natural image generation provides new insights into both adversarial robustness and generative modeling. The same principles that make representations robust against attack—consistency across scales and alignment with natural image statistics—also enable high-quality generation from discriminative models without additional training.

4. Method

4.1. From Discrimination to Generation

Every discriminative model contains within it the seeds of a generative model – the challenge lies in accessing

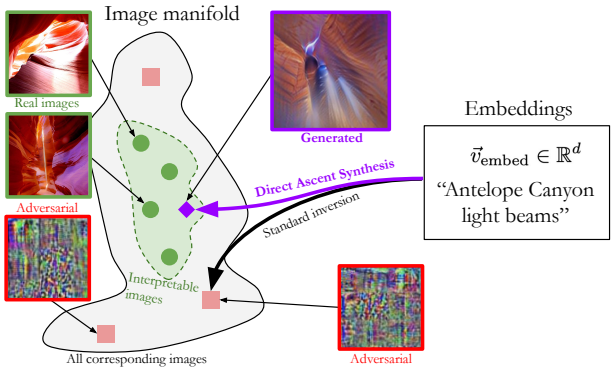

Figure 5. Mapping between images and embeddings. A region of all images corresponding to a {text, image} embedding contains interpretable images as well as noise-like adversarial patterns. Reconstructing an image from an embedding typically leads to such a degenerate noisy image. With Direct Ascent Synthesis, the reconstructed image lands among interpretable images within the manifold by default.

these capabilities effectively. Models like CLIP map images $I \in \mathbb{R}^{H\times W\times C}$ to embedding vectors $v \in \mathbb{R}^d$, learning rich representations that capture both semantic content and natural image structure. While the forward mapping ($f : I \to v$) is straightforward, the reverse mapping ($f^{-1} : v \to I$) has traditionally been seen as problematic due to its one-to-many nature and tendency to produce adversarial patterns (see Figure 5 for a diagram capturing this). The dimensionality of such manifolds was studied in e.g. (Fort et al., 2022).

Our key insight is that this perceived limitation is actually an opportunity: the space of possible inversions contains both natural images and adversarial patterns (as both are genuinely predictive of the embedding (Ilyas et al., 2019)), and careful optimization can guide us toward the former. Given a target embedding u (e.g., from a text description), we measure alignment through cosine similarity:

$$\text{score}(I) = \cos(f(I), u) \quad (3)$$

4.2. Multi-Resolution Optimization

The critical innovation in DAS is decomposing the optimization across multiple scales – a choice that proves surprisingly powerful in guiding solutions toward natural images. We break the degeneracy by decomposing the optimization across multiple scales. Instead of directly optimizing pixels, we express the image as a sum of resolution components:

$$I = \frac{1}{2} + \frac{1}{2}\tanh\left(\sum_{r\in\rho} \text{resize}_{224}(P_r)\right) \quad (4)$$

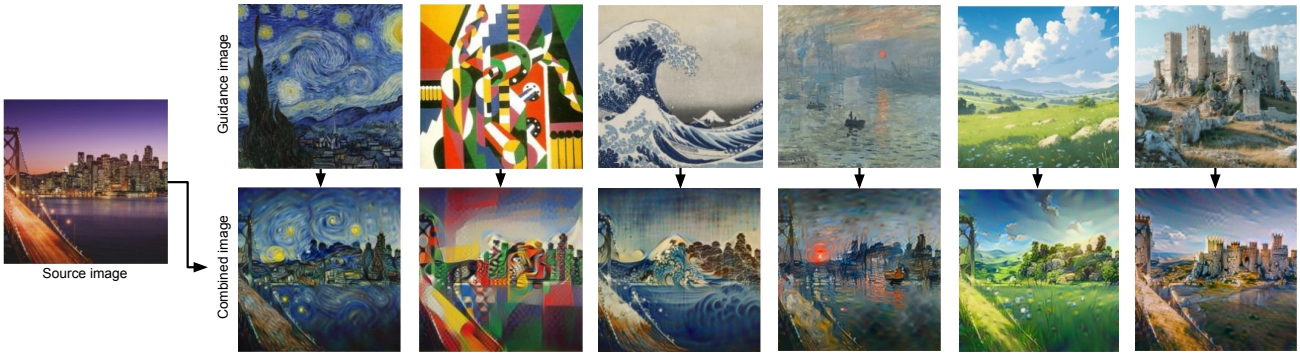

Figure 6. Direct Ascent Synthesis enables efficient neural style transfer without the artifacts common in pixel-space optimization. Starting from a source image and guidance image, we are able to effectively combine the two using DAS. This demonstrates that our multi-resolution framework naturally extends beyond CLIP-guided generation to other optimization-based image synthesis tasks.

where $P_r \in \mathbb{R}^{r\times r\times 3}$ represents the image component at resolution r, and ρ spans from 1×1 to 224×224. The $\tanh$ transformation maps unbounded optimization values to valid pixel intensities while maintaining gradient flow.

The optimization objective becomes:

$$\sum_{i,j} \frac{\partial \text{score}_i(\text{augment}_j(I(P_1,\ldots,P_{224})))}{\partial(P_1,\ldots,P_{224})} \tag{5}$$

where i indexes multiple CLIP models and j indexes augmentations. This formulation has several key properties: 1) Components are optimized simultaneously across all resolutions, 2) Gradients naturally distribute across scales based on their importance, 3) High-frequency adversarial patterns are suppressed by scale decomposition.

The resulting resolution components of a generated image are shown in Figure 2, together with the power spectrum of the generated image, which follows a $1/f^2$ distribution (slope ≈ -2), characteristic of natural images (Ruderman, 1994; Hyvärinen et al., 2009). This demonstrates that our multi-resolution prior effectively guides optimization toward perceptually valid solutions.

4.3. Implementation Details

We employ several techniques to ensure stable and high-quality generation:

Augmentation. Two minimal augmentations prove crucial: random x-y shifts and pixel noise. These work in synergy with the multi-resolution prior – neither is sufficient alone but together they enable robust generation. More complex augmentations might realistically lead to higher-quality generation.

Shift Handling. Rather than traditional padding approaches, we generate images at $(H + 2s) \times (W + 2s)$ resolution where s is the maximum shift. This provides a natural buffer for shift augmentation, with the final image center-cropped to $H \times W$. Individual x-y shifts are guaranteed never to exceed the larger image, making sure padding is not necessary.

Model Ensemble. We average gradients across three CLIP models: OpenAI ViT-B/32 and two OpenCLIP (Cherti et al., 2023) ViT-B/32 variants trained on different datasets. This (marginally) improves generation quality, however, a single model is sufficient. We found, however, that some CLIP models were particularly bad at being turned into generators without any obvious reason why.

4.4. Extensions

The framework naturally supports several useful extensions:

Multiple Target Vectors. Generation can be guided by multiple weighted targets: $\sum_i w_i \text{score}(v, u_i)$. This enables fine control through prompt combinations (e.g., enhancing aesthetics with "volumetric lighting" while suppressing text with "Optical Character Recognition"). We use the latter to prevent CLIP from spelling out the semantic content of the desired generation.

Reference Images Target embeddings can come from either text or reference images ($f(I_{\text{ref}}) = u$), enabling style transfer and reconstruction tasks. Despite CLIP's compression from 150,528 dimensions to 512, reconstruction often preserves both semantic content and style elements. See Figure 6 for style transfer and Figure 9 for reconstruction examples.

(a) Volcanic eruption in Iceland

(b) Cambridge UK + winter night

Figure 7. Four independent generations of "a photo of a volcanic eruption in Iceland" on the left, and "a beautiful photo of Cambridge UK, detailed" with an additional prompt of "winter night". The generations used 3 CLIP models at once, and a corrective prompt of "Optical Character Recognition" with a weight of -0.6 to avoid text appearing in the images. The four samples demonstrate generation diversity.

5. Experiments and Analysis

We evaluate Direct Ascent Synthesis through a comprehensive set of experiments designed to probe both its generative capabilities and its relationship to discriminative representations. Our analysis focuses on four key aspects of the method: generation consistency, controlled modification, reconstruction fidelity, and versatility across different applications.

The specific optimization details were kept as simple as possible for the sake of understanding the underlying generative capabilities of DAS. We optimized for 100 steps with Stochastic Gradient Descent (Robbins & Monro, 1951) at the learning rate of 2×10^{-1}. The added noise had a standard deviation of 0.2, the x-y plane jitter in the ± 56 range (implying a generation of a 336×336 from which we then center-cropped the 224×224 in the middle), and using 32 augmentations at once in a batch. We also used 3 CLIP models in an ensemble: OpenAI ViT-B/32 and two OpenCLIP ViT-B/32. All models we used are based on the Vision Transformer architecture (Dosovitskiy et al., 2021), however, we have verified that non-ViT models work similarly well.

Figure 3 shows 18 images generated using DAS from a diverse set of prompts.

5.1. Generation Quality and Consistency

A fundamental question for any generative method is whether it can consistently produce coherent results. Figure 7 demonstrates DAS's reliability across multiple runs on two challenging prompts: a dynamic natural phenomenon (volcanic eruption) and a complex architectural scene (Cambridge on a winter night). These examples were chosen specifically to test the method's ability to handle both natural and man-made structures, dynamic events, and specific lighting conditions.

Our analysis reveals three key properties of the generation process:

- **Semantic Consistency:** Each set of generations maintains consistent high-level features while varying in specific details. For the volcanic scenes, this manifests as consistent plume structure and landscape integration. In the Cambridge scenes, we observe reliable architectural motifs and winter atmosphere, suggesting that the optimization reliably finds meaningful regions in CLIP's representation space.

- **Compositional Understanding:** The images demonstrate sophisticated composition without explicit training. The volcanic scenes balance foreground drama with environmental context, while the Cambridge scenes show an understanding of architectural perspective and nighttime illumination. This suggests that our multi-resolution optimization effectively accesses CLIP's learned understanding of scene structure.

- **Natural Variation:** The differences between runs exhibit variations characteristic of natural images—lighting changes, slight perspective shifts, and detail variations—rather than adversarial patterns. This indicates that our multi-resolution prior successfully constrains the optimization to the natural image manifold.

5.2. Controlled Modification

Figure 8 explores DAS's capacity for targeted image modification, demonstrating both local adjustments (surface changes) and global transformations (environmental shifts). These experiments reveal several important capabilities:

- **Structure Preservation:** Core scene geometry and spatial relationships persist across transformations, indicating that our optimization respects structural features encoded in CLIP's representation space.

- **Semantic Control:** The modifications show precise response to textual prompts while maintaining physical plausibility. Snow accumulates naturally on surfaces, the volcano emerges with appropriate atmospheric effects, and surface textures change coherently.

(a) Original

(b) "snowy road"

(c) "a volcano eruption"

(d) "asphalt road"

Figure 8. Starting from the original image, we run generation towards the specified prompt. The modifications demonstrate both local changes (road surface, volcano in the background) and global scene transformations (winter and snow) while maintaining spatial coherence.

- **Multi-Scale Coordination:** New elements integrate seamlessly across different spatial scales. This is particularly evident in the volcano example, where both large-scale landscape changes and local atmospheric effects are coordinated.

5.3. Embedding-Guided Reconstruction

Image reconstruction from CLIP embeddings provides a particularly rigorous test of our method, as it requires recovering high-dimensional image structure from highly compressed representations (from 150,528 dimensions to just 512). Figure 9 demonstrates that DAS can recover substantial semantic and stylistic information, with the reconstructions showing:

- **Semantic Preservation:** Major scene elements and their relationships are consistently recovered
- **Style Retention:** Color schemes, lighting conditions, and artistic qualities transfer effectively
- **Compositional Fidelity:** Overall layout and spatial organization remain intact
- **Natural Variation:** Fine details vary while maintaining scene coherence

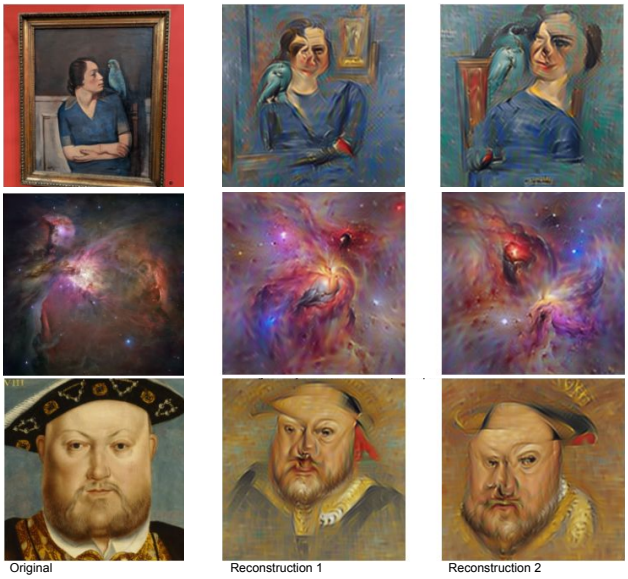

Figure 9. Reconstructing an image from its embedding. Instead of a text prompt, we used an embedded original image to guide the Direct Ascent Synthesis generation. Two resulting reconstructions are shown for each image, demonstrating consistent recovery of major semantic elements and style while allowing natural variations in specific details. Given the 300:1 dimensionality reduction from an image to an embedding, the recovery is impressive.

This performance is particularly noteworthy given that CLIP was never trained for reconstruction or compression tasks. Yet it can recover major aspects of an image from its 300:1 compressed embedding.

5.4. Specialized Applications

To explore the versatility of our framework, we tested DAS on specialized generation tasks that typically require dedicated solutions. Figure 10 shows the generation of national flags, a task requiring precise geometric patterns and symbolic elements. The results demonstrate that DAS can handle both rigid geometric constraints and subtle style elements, like the precise proportions of the Swiss cross and the complex star pattern of the Brazilian flag. While the generation is far from perfect, the flags are clearly recognizable.

Figure 11 showcases inpainting capabilities, where DAS must generate content that seamlessly integrates with existing image context. The successful completion of the city skyline demonstrates that our multi-resolution optimization naturally handles boundary conditions and structural continuity without additional constraints.

These specialized applications highlight a key advantage of our approach: a single optimization framework can ad-

(a) Japan (b) Brazil (c) UN (d) Swiss

Figure 10. Generating flags demonstrates DAS's ability to handle precise geometric patterns and symbolic elements. The prompts combine "the flag of [X]" with structural guidance (-0.3 × "Optical Character Recognition", 0.6 × "cohesive single subject", -0.3 × "multiple exposure").

Figure 11. An example of inpainting using Direct Ascent Synthesis. The masked image was filled in using the prompt "a city skyline at night", demonstrating seamless integration of generated content with existing context.

dress diverse synthesis tasks without task-specific training or architectural modifications. This versatility stems from the rich representational knowledge captured by CLIP combined with the natural structure-preserving properties of multi-resolution optimization.

5.5. Style Transfer

We can modify a starting image toward the embedding of a "style" image easily using DAS. This functions effectively as a natural version of style transfer. See Figure 6 for details. The resulting generations respect the structure of the original image while copying the style and local content from the guiding image.

Beyond CLIP, we find DAS a useful drop-in replacement for other techniques that rely on pixel-space optimization, such as the more traditional style transfer approaches. For example, following (Gatys et al., 2016) we compare the results on raw pixels with those obtained by using our method and, while the comparison is somewhat subjective, find that the latter tends to produce more pleasing results with less high-frequency artifacts and in fewer steps. See Figure 12.

6. Discussion and Future Work

Our results with Direct Ascent Synthesis reveal fundamental insights about the relationship between discrimination and generation in deep neural networks. By demonstrating that discriminative models contain rich generative knowledge that can be accessed through careful optimization, we challenge several conventional assumptions in the field.

(a) Style (b) Content (c) Pixels (d) DAS

Figure 12. Applying style transfer while optimizing raw pixels (c) vs DAS (d).

6.1. Theoretical Implications

The success of DAS suggests that the traditional division between discriminative and generative models may be more fluid than previously thought. It appears that seeds of generation are hidden in all discriminative models, and we just needed a better way of eliciting them. Several key insights emerge:

- **Representation Unification:** The ability of discriminative models to support high-quality generation suggests a fundamental unity in neural representations. Rather than learning strictly task-specific features, these models appear to capture a more complete understanding of visual structure that can support both discrimination and generation. This challenges the traditional view of separate representational requirements for these tasks.

- **Information Preservation:** Our results indicate that discriminative training, despite optimizing only for classification or similarity metrics, preserves much of the information needed for generation. This suggests that the process of learning to discriminate naturally encodes generative capabilities as a byproduct, pointing to deeper connections between these two aspects of visual processing.

- **Optimization vs. Architecture:** DAS demonstrates that the key to accessing generative capabilities may lie more in the optimization process than in network architecture. This suggests that the historical focus on architectural differences between discriminative and generative models may have overshadowed the importance of how we access and utilize their learned representations.

6.2. Practical Implications

Beyond theoretical insights, DAS has several important practical implications:

- **Resource Efficiency:** By leveraging pretrained discriminative models, DAS enables image generation with significantly lower computational requirements than traditional generative approaches.
- **Architecture Design:** The success of multi-resolution optimization suggests new directions for neural architecture design that explicitly incorporate scale-space structure.
- **Model Reuse:** DAS demonstrates that existing discriminative models may have untapped capabilities that can be accessed through novel optimization strategies.

6.3. Connections to Model Interpretability

Our work with DAS has significant implications for model interpretability research. The ability to extract coherent generative capabilities from discriminative models suggests that standard interpretability approaches may overlook important model properties. While traditional interpretability tools focus on analyzing individual neurons or attention patterns, DAS reveals emergent capabilities that arise from the interaction of model components across different scales.

The success of our multi-resolution optimization approach provides evidence that model representations are naturally organized hierarchically, aligning with recent work in circuit-style interpretability (Elhage et al., 2022; Olah et al., 2020). This suggests that models may learn to decompose visual information across multiple scales not just for discrimination tasks, but as a fundamental organizational principle that supports both discriminative and generative capabilities.

Perhaps most intriguingly, the ability to extract high-quality generative behavior from discriminative models challenges the traditional dichotomy between discriminative and generative architectures from an interpretability perspective. This implies that the apparent distinction between these model types may be more a function of how we access their capabilities than of fundamental differences in their learned representations (Bau et al., 2019a).

Beyond its primary application as a synthesis method, DAS can be viewed as a novel interpretability technique. By revealing what generative information is preserved in discriminative models, it provides a new lens for understanding what these models actually learn. This suggests that the space of possible interpretability tools may be much richer than previously recognized, particularly when we consider emergent capabilities that span traditional architectural boundaries.

6.4. Limitations and Open Questions

While DAS achieves impressive results, several important challenges remain: Generation quality can vary across runs and prompts, suggesting room for improving optimization stability. While empirically effective, we lack a complete theoretical framework (that is emerging for diffusion models, e.g. Kamb & Ganguli (2024)) explaining why multi-resolution optimization so effectively prevents adversarial solutions.

Our results suggest an intriguing connection between adversarial robustness and generative capabilities. The same multi-resolution structure that enables coherent generation also appears to prevent adversarial patterns, suggesting that natural image statistics may play a crucial role in both phenomena. This raises the possibility that advances in understanding one area could inform the other – perhaps robust models are inherently better at generation, or generative capabilities could be used as a proxy for robustness. (Fort (2025) provides early indications that adaptive adversarial attacks against a strong multi-resolution model often produce human-interpretable changes to the image.)

6.5. Future Directions

Our work opens several promising avenues for future research:

Unified Training Objectives. Future work could explore training objectives that explicitly optimize for both discriminative and generative capabilities, potentially leading to more efficient and versatile models. This might involve novel loss functions that balance feature discrimination with generative consistency.

Theoretical Frameworks. Developing formal mathematical frameworks that unify discriminative and generative learning could provide deeper insights into why methods like DAS work. This might draw on information theory, optimal transport, or other theoretical tools to characterize the relationship between these traditionally separate paradigms.

Cross-Domain Applications. The principles underlying DAS might extend beyond vision to other domains where discriminative and generative tasks have traditionally been separated, such as natural language processing or audio synthesis. This could lead to new training-free generation methods across multiple modalities.

DAS + explicit generative training. Given that DAS can elicit generative capabilities from discriminative models, it would be intriguing to explore whether its generation can be further improved by additional training explicitly geared towards generation. Merging DAS with diffusion models

might be such an avenue.

Using intermediate layers. In DAS we have been using the final layer embedding to guide the generation process. In line with the self-ensemble approach in Fort & Lakshminarayanan (2024), non-final layers could be used as well, providing a more detailed control over the generation process on different levels of abstraction.

7. Conclusion

We have presented Direct Ascent Synthesis, demonstrating that high-quality image generation is possible through direct optimization of discriminative model representations. This finding challenges conventional assumptions about the necessity of dedicated generative training and suggests new directions for understanding and advancing visual synthesis. Our work reveals a deep connection between model inversion, adversarial examples, and image generation—problems that have traditionally been studied separately but share fundamental mathematical characteristics.

The key insight of DAS is that the challenges of model inversion, which have primarily been viewed through the lens of adversarial attacks, can be transformed into opportunities for synthesis through careful regularization. Where adversarial attacks exploit the degeneracy of the inversion problem to find perceptually misleading solutions, our multi-resolution approach harnesses this same flexibility to find natural, semantically meaningful images. This suggests that the perceived limitations of discriminative models—their vulnerability to adversarial examples and the apparent difficulty of inverting their representations—may actually reflect unexploited generative capabilities.

The success of our simple approach raises fundamental questions about the nature of visual representation in neural networks. Perhaps the sharp distinction between discriminative and generative models has been somewhat artificial—both types of models may be learning similar underlying representations, just accessed in different ways. The fact that adversarial patterns and natural images can arise from the same optimization process, differentiated only by their scale-space structure, hints at deep connections between robustness, generalization, and generation in neural networks. It appears that seeds of generation are hidden in all discriminative models.

Looking forward, this work suggests that the boundaries between discrimination, generation, and robustness may be more fluid than previously recognized. By viewing these challenges through a unified lens of representation and optimization, we may discover new approaches that simultaneously advance all three areas. DAS represents an important step in this direction, demonstrating that with appropriate optimization techniques, discriminative models can transcend their traditional role and serve as powerful tools for image synthesis.

References

Bau, D., Zhu, J.-Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., and Torralba, A. Gan dissection: Visualizing and understanding generative adversarial networks. In *International Conference on Learning Representations*, 2019a.

Bau, D., Zhu, J.-Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., and Torralba, A. Seeing what a gan cannot generate, 2019b. URL https://arxiv.org/abs/1910.11626.

Burt, P. and Adelson, E. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31 (4):532–540, 1983. doi: 10.1109/TCOM.1983.1095851.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, June 2023. doi: 10.1109/cvpr52729.2023.00276. URL http://dx.doi.org/10.1109/CVPR52729.2023.00276.

Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., and Raff, E. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pp. 88–105. Springer, 2022.

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A., Caron, M., Geirhos, R., Alabdulmohsin, I., Jenatton, R., Beyer, L., Tschannen, M., Arnab, A., Wang, X., Riquelme, C., Minderer, M., Puigcerver, J., Evci, U., Kumar, M., van Steenkiste, S., Elsayed, G. F., Mahendran, A., Yu, F., Oliver, A., Huot, F., Bastings, J., Collier, M. P., Gritsenko, A., Birodkar, V., Vasconcelos, C., Tay, Y., Mensink, T., Kolesnikov, A., Pavetić, F., Tran, D., Kipf, T., Lučić, M., Zhai, X., Keysers, D., Harmsen, J., and Houlsby, N. Scaling vision transformers to 22 billion parameters, 2023. URL https://arxiv.org/abs/2302.05442.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Ndousse, K., Amodei, D., Jones, A., DasSarma, N., Askell, D., Wang, P., and Chen, A. A mathematical framework for transformer circuits. *Anthropic Technical Report*, 2022.

Fort, S. Adversarial examples for the openai clip in its zero-shot classification regime and their semantic generalization, Jan 2021a. URL https://stanislavfort.github.io/2021/01/12/OpenAI_CLIP_adversarial_examples.html.

Fort, S. Pixels still beat text: Attacking the openai clip model with text patches and adversarial pixel perturbations, March 2021b. URL https://stanislavfort.github.io/2021/03/05/OpenAI_CLIP_stickers_and_adversarial_examples.html.

Fort, S. Adversarial vulnerability of powerful near out-of-distribution detection, 2022. URL https://arxiv.org/abs/2201.07012.

Fort, S. A note on implementation errors in recent adaptive attacks against multi-resolution self-ensembles, 2025. URL https://arxiv.org/abs/2501.14496

Fort, S. and Lakshminarayanan, B. Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness, 2024. URL https://arxiv.org/abs/2408.05446.

Fort, S., Ren, J., and Lakshminarayanan, B. Exploring the limits of out-of-distribution detection, 2021. URL https://arxiv.org/abs/2106.03004.

Fort, S., Cubuk, E. D., Ganguli, S., and Schoenholz, S. S. What does a deep neural network confidently perceive? the effective dimension of high certainty class manifolds and their low confidence boundaries, 2022. URL https://arxiv.org/abs/2210.05546.

Gatys, L. A., Ecker, A. S., and Bethge, M. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Hyvärinen, A., Hurri, J., and Hoyer, P. O. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*, volume 39 of *Computational Imaging and Vision*. Springer, 2009. ISBN 978-1-84882-491-1. doi: 10.1007/978-1-84882-491-1.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features, 2019. URL https://arxiv.org/abs/1905.02175.

Jeantet, C., Caharel, S., Schwan, R., Lighezzolo-Alnot, J., and Laprevote, V. Factors influencing spatial frequency extraction in faces: A review. *Neuroscience and Biobehavioral Reviews*, 93:123–138, 2018. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2018.03.006. URL https://www.sciencedirect.com/science/article/pii/S014976341730204X.

Kamb, M. and Ganguli, S. An analytic theory of creativity in convolutional diffusion models, 2024. URL https://arxiv.org/abs/2412.20292.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Lindeberg, T. Scale-space theory: a basic tool for analyzing structures at different scales. *Journal of Applied Statistics*, 21(1-2):225–270, 1994. doi: 10.1080/757582976. URL https://doi.org/10.1080/757582976.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks, 2019. URL https://arxiv.org/abs/1706.06083.

Mahendran, A. and Vedaldi, A. Understanding deep image representations by inverting them, 2014.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, 2016. URL https://arxiv.org/abs/1605.09304.

Olah, C., Mordvintsev, A., and Schubert, L. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214%2Faoms%2F1177729586.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Ruderman, D. L. Statistics of natural images: Scaling analysis and the scale-space paradigm. *Physical Review Letters*, 73(6):814–817, 1994. doi: 10.1103/PhysRevLett.73.814.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

van der Schaaf, A. and van Hateren, J. Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 36(17):2759–2770, September 1996. ISSN 0042-6989. Relation: http://www.rug.nl/informatica/organisatie/overorganisatie/iwi Rights: University of Groningen. Research Institute for Mathematics and Computing Science (IWI).

Whitaker, J. imstack: Image stack exploration and analysis. https://johnowhitaker.github.io/imstack/, 2022. Accessed: 2024-01-31.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. Understanding neural networks through deep visualization, 2015. URL https://arxiv.org/abs/1506.06579.