Blending Art and Intelligence: Advances in Neural Style Transfer and Image Synthesis

Preprint · December 2023

DOI: 10.13140/RG.2.2.12000.60160

Grace Scott
University of Minnesota

Blending Art and Intelligence: Advances in Neural Style Transfer and Image Synthesis

Grace Scott, Chloe Green, Brooke Baker

1. Introduction

The convergence of art and artificial intelligence has led to groundbreaking advancements in the field of computer vision, where techniques such as Neural Style Transfer and Image Synthesis are redefining creative expression. These innovations not only showcase the immense potential of AI in interpreting and generating visuals but also open up unprecedented opportunities for creativity, design, and technology. By blending computational precision with artistic ingenuity, these methods allow AI to create visuals that are both meaningful and transformative, pushing the boundaries of how we perceive and produce art.

Neural Style Transfer, in particular, has emerged as a captivating technique that blends the artistic essence of one image with the structural content of another. Imagine transforming a simple photograph into a vivid masterpiece inspired by Van Gogh or Picasso, or reimagining everyday visuals with the texture and colors of Renaissance paintings. This process leverages Convolutional Neural Networks (CNNs) to disentangle and recombine the stylistic and structural elements of images. By defining a loss function that balances content preservation with stylistic adaptation, Neural Style Transfer enables the creation of artwork that bridges the gap between classical art forms and modern computational methods.

On the other hand, Image Synthesis extends beyond transforming existing visuals—it generates entirely new images from scratch. Powered by Generative Adversarial Networks (GANs), this technique introduces an adversarial framework where a generator creates synthetic images and a discriminator evaluates their authenticity. Through this iterative process, GANs produce hyper-realistic visuals that have applications spanning art, commerce, data augmentation, and beyond. Together, Neural Style Transfer and Image Synthesis represent a transformative intersection of art, technology, and artificial intelligence, offering tools to redefine creative workflows and enrich human expression.

Neural Style Transfer: Transforming Artistic Imagination Neural Style Transfer (NST) revolutionizes the relationship between technology and creativity by enabling the seamless blending of artistic style and content. It operates by separating and reassembling the fundamental aspects of images—content, which defines the structure or layout, and style, which encompasses color, texture, and patterns. Using pre-trained Convolutional Neural Networks (CNNs) such as VGG-19, NST extracts features from both content and style images to compute losses that guide the generation of a new, stylized image.

The process involves three main inputs: a content image, a style image, and a generated image. The content loss ensures that the structure of the content image is preserved, while the style loss enforces the application of textures and patterns from the style image. Through iterative optimization using gradient descent, the generated image evolves until the combined loss is minimized, resulting in a harmonious blend of content and style. Over time, advancements like Adaptive Instance Normalization (AdaIN) and feed-forward networks have significantly enhanced the efficiency and flexibility of NST, enabling real-time style transfer for applications in live media and interactive experiences.

Image Synthesis: Generating Reality from Randomness While Neural Style Transfer transforms existing visuals, Image Synthesis takes a bold step by crafting entirely new images from random noise. This is achieved through the adversarial framework of GANs, where two neural networks—the generator and the discriminator—engage in a dynamic interplay. The generator's goal is to create images that can fool the discriminator into classifying them as real, while the discriminator's objective is to distinguish between genuine and synthetic images. This competition drives the generator to produce increasingly realistic outputs over successive iterations.

GANs have revolutionized fields such as art and commerce, where they enable the creation of unique digital artwork and design prototypes. In data augmentation, GANs address the challenge of limited training datasets by generating diverse and realistic samples to enhance model performance. They are also widely used in tasks like image-to-image translation, where one type of image is transformed into another—for example, converting sketches into fully colored illustrations or daytime photos into nighttime scenes. The advent of models like StyleGAN and Progressive GANs has further refined the quality, control, and resolution of synthesized images, solidifying GANs as a cornerstone of modern AI.

Merging Neural Style Transfer and Image Synthesis The combination of Neural Style Transfer and Image Synthesis unlocks even greater creative possibilities. By generating base images using GANs and applying Neural Style Transfer, it becomes possible to produce visuals that are not only realistic but also rich in artistic expression. This hybrid approach has applications in artistic image generation, where stylized visuals are created from scratch, and enhanced data augmentation, where synthetic images are diversified to train machine learning models.

The intersection of these two techniques also offers transformative tools for artists and designers, enabling them to seamlessly generate and stylize visuals. For example, GANs can create detailed image layouts, which can then be infused with the stylistic characteristics of renowned art movements using NST. This synergy between generative models and style transfer represents a future where creativity is amplified by AI, empowering individuals to explore new artistic frontiers.

The Broader Implications As these techniques continue to evolve, their influence extends far beyond art and design. Neural Style Transfer and Image Synthesis are reshaping industries such as film, animation, healthcare, and education. Filmmakers can stylize entire scenes to evoke specific moods or historical aesthetics, while medical professionals can use high-resolution synthesized images to visualize complex biological processes. Additionally, educators can create engaging visual aids to enhance student comprehension in subjects ranging from biology to geography.

With advancements in real-time processing and multimodal synthesis, the future holds immense potential for integrating these methods into immersive multimedia experiences. From virtual reality and interactive storytelling to AI-assisted creative tools, Neural Style Transfer and Image Synthesis are paving the way for a new era of human-AI collaboration, where technology serves as both a partner and an enabler of artistic innovation.

2. Related Work

The Rise of Unsupervised Image Synthesis [51, 8, 39, 81, 82]. Given the constraints of paired data, unsupervised image synthesis methods have emerged as a promising alternative. Techniques like CycleGAN [98], DualGAN [32], and DiscoGAN [75] have gained widespread attention for their ability to learn mappings between domains without requiring paired examples.

Unsupervised approaches typically leverage large collections of images from each domain and use cycle-consistency constraints to ensure that generated outputs remain coherent. For instance, translating an image from domain A to domain B and back to domain A should yield the original input image, thereby enforcing consistency between the domains.

While unsupervised methods have demonstrated impressive results, their performance often depends on the availability of abundant data and may struggle in scenarios where domain-specific details require more explicit guidance or labeling.

Semi-Supervised Approaches to Image Synthesis To address the limitations of fully unsupervised methods, researchers have explored semi-supervised learning approaches for image synthesis. Semi-supervised techniques combine a small amount of paired data with a larger pool of unpaired data to improve the quality of generated images. These methods can be particularly useful in scenarios like old movie restoration [52] or genomics applications [62], where limited expert annotations can provide critical guidance.

Prominent semi-supervised learning methods such as those proposed by Kingma et al. [33], Rasmus et al. [59], Berthelot et al. [7], and Zhang et al. [89] have demonstrated the potential of this hybrid approach. By leveraging small amounts of aligned data alongside larger unpaired datasets, semi-supervised models achieve more compelling results compared to their unsupervised counterparts while significantly reducing the reliance on fully paired datasets.

Few-Shot and One-Shot Learning in Image Synthesis In contrast to traditional machine learning models, which require extensive training data, humans can learn new tasks from just a few examples by leveraging prior knowledge and experiences. Inspired by this capability, few-shot and one-shot learning approaches have been developed to enable neural networks to generalize effectively with limited data.

Meta-learning and few-shot learning frameworks [93, 63] mimic this human ability by incorporating mechanisms for knowledge transfer and generalization. For instance, techniques such as TuiGAN [37], Few-Shot GAN [41], and Zero-Shot Transfer GAN (ZSTGAN) [38] aim to translate between domains with only a few or even a single example from the target domain. These models are particularly valuable in applications where data collection is constrained or expensive.

Multimodal Image Synthesis. Most image translation techniques focus on one-to-one mappings, generating a single output for a given input. However, real-world translation tasks are often inherently ambiguous, as one input image can correspond to multiple plausible outputs. Multimodal image synthesis addresses this challenge by learning a distribution of potential outputs in the target domain, ensuring that generated images are diverse yet consistent with the input image.

Multimodal approaches produce outputs that reflect different variations while preserving the essential characteris-

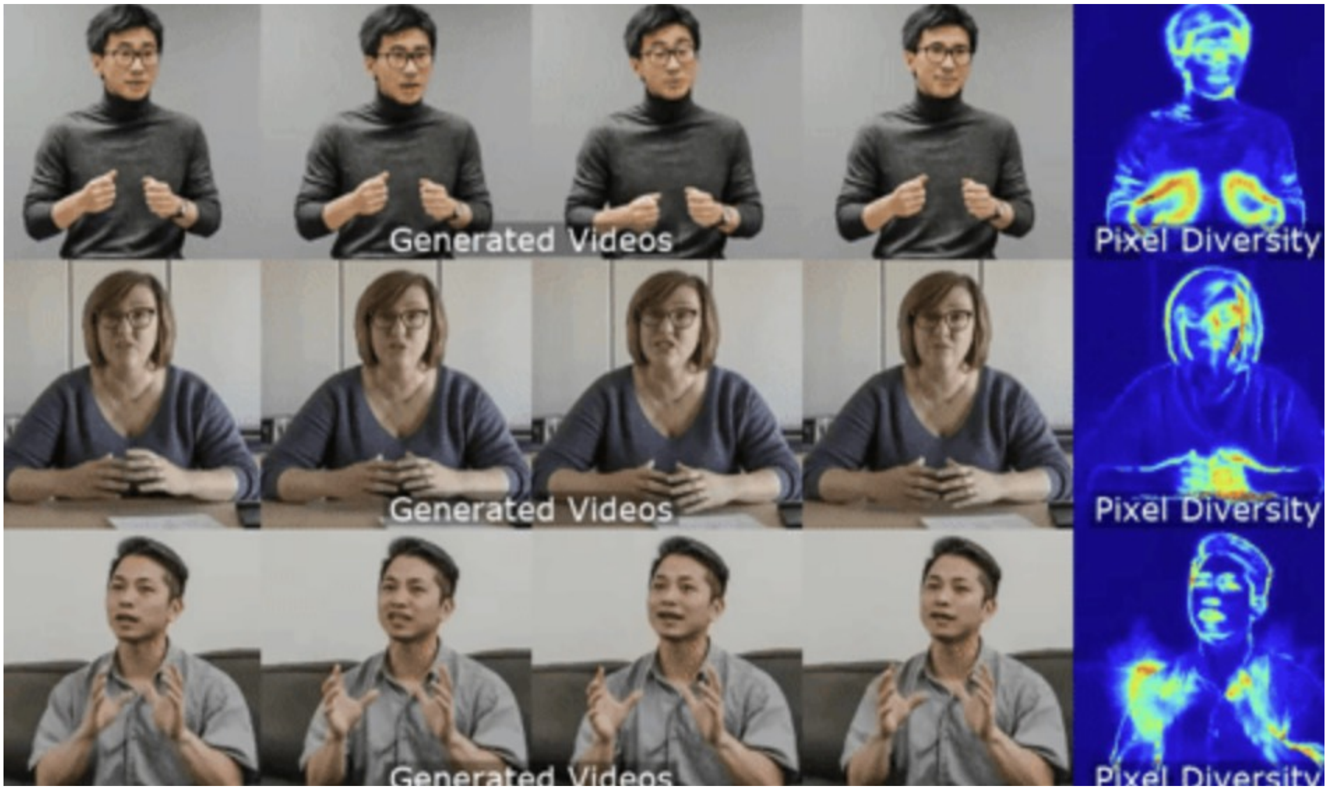


Figure 1.

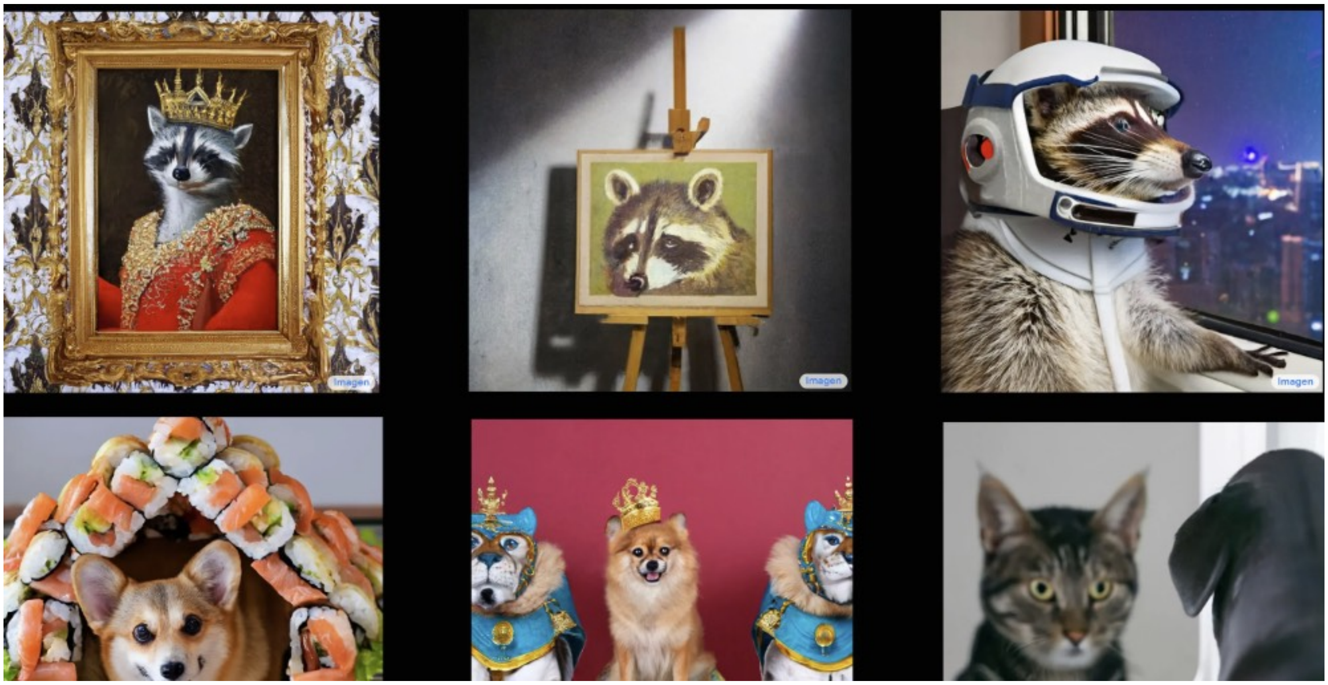
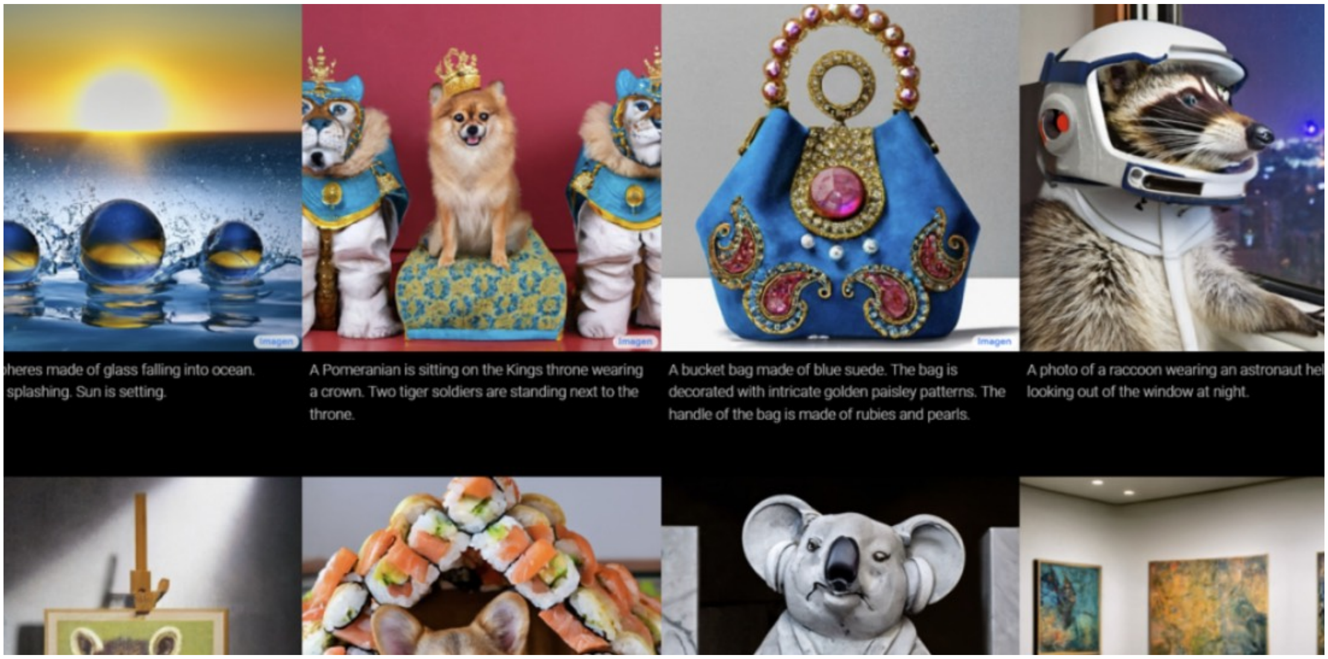


Figure 2.

tics of the source image. This diversity is critical in applications like style transfer, where multiple artistic interpretations of a photograph might be desired, or in medical imaging, where variability can provide additional diagnos-



heres made of glass falling into ocean. splashing. Sun is setting.

A Pomeranian is sitting on the Kings throne wearing a crown. Two tiger soldiers are standing next to the throne.

A bucket bag made of blue suede. The bag is decorated with intricate golden paisley patterns. The handle of the bag is made of rubies and pearls.

A photo of a raccoon wearing an astronaut hel looking out of the window at night.

Figure 3.

tic insights.

The evolution of controllable image synthesis techniques—from supervised to unsupervised, semi-supervised, and few-shot learning—has significantly expanded the scope and applicability of generative models. While early methods relied on paired datasets, modern approaches have demonstrated the ability to operate under more flexible and challenging settings, including unpaired and few-shot scenarios. Multimodal image synthesis further enhances the creative and practical potential of these models by offering diverse outputs that remain faithful to the input.

As the field continues to progress, the integration of these methodologies with advanced neural architectures promises to unlock new possibilities in image synthesis, bridging the gap between technical innovation and real-world application.

Most of computer visions problems can be seen as an image-to-image translation problem, mapping an image from one domain to another image in different domain. As an illustration, super-resolution can be viewed as a concern of mapping a low-resolution image to a similar high-resolution one; image colorization is a problem of mapping a gray-scale image to a corresponding color one. The problem can be investigated in supervised and unsupervised learning methods. In the supervised approaches, paired of images in various domains are available [29]. In the unsupervised models, only two separated sets of images are available in which one composed of images in one domain and the other composed of different domain images—there is no paired samples representing how an image can possibly translated to a corresponding image in different domain. For lack of corresponding images, the unsupervised image-to-image translation problem is considered more difficult, but it is more feasible because training data collection is easier.

When assessing the image translation problem from a likelihood viewpoint, the main challenge is to learn a mutual distribution of images in different domains. In the unsupervised setting, the two sets composed of images from two minor distributions of different domains, and the task is to gather the cooperative distribution by utilizing these images. However, driving the joint distribution from the minor distributions is extremely ill-posed problem. In this section, we discuss the image-to-image translation methods. Image-to-image translation is similar to style transfer, which as the input receives a style image and a content image. The model output is an image that has the content of the content image and the style of the style image. It is not only transferring the images' styles, but also manipulates features of objects. This section lists several models that are proposed for image-to-image translation from supervised methods to unsupervised ones.

2.1. Supervised Translation

Isola et al. [29] made significant advancements in the field of image-to-image translation by proposing a method that integrates adversarial network losses with L_1 regularization. This approach not only trains the generator to pass the discriminator's filtering process but also ensures that the generated images are realistic and closely resemble the ground-truth images. The use of L_1 loss, as opposed to L_2, results in less blurry images, which was the primary motivation behind its adoption. Here, $x, y \sim p(x, y)$ represents images with different styles but from the same scene, while $z \sim p(z)$ denotes random noise. where the hyperparameter λ balances the two loss functions. Isola et al. [29] observed that the noise z had minimal influence on the results. Instead, they suggested using dropout during both training and testing as a substitute for random noise.

The generator G in their model was based on a modified U-Net structure, incorporating multi-scale connections to link encoder layers to corresponding decoder layers, thereby sharing low-level information like object edges. Additionally, the authors introduced the `PatchGAN` discriminator, which focuses on classifying $N \times N$ image patches rather than the entire image. This strategy ensures the discriminator captures high-frequency details by concentrating on local patches, which proved sufficient for achieving fine-grained outputs.

Supervised and Semi-Supervised Approaches in Image-to-Image Translation. Yoo et al. introduced a supervised algorithm for image-to-image translation with a secondary discriminator, D_{pair}, which evaluates whether pairs of images from multiple domains are related.

In this framework, X_s represents the source domain image, X_t the target domain ground-truth image, and $X_{\bar{t}}$ an irrelevant image from the target domain. The generator transfers X_s to $\hat{X}_t$, its corresponding image in the target domain.

Zareapoor et al. extended adversarial networks to semi-supervised settings, proposing models for dataset balancing in mechanical devices. These approaches demonstrated high accuracy and efficiency by integrating multi-instance learning for human pose estimation. Shamsolmoali et al. tackled imbalanced class problems using capsule adversarial networks with minority class augmentation. Zhang et al. proposed the DRCW-ASEG method, which generates synthetic examples to address multi-class imbalanced problems, achieving improved classification accuracy.

Advances in Pix2Pix and Beyond. The pix2pix framework introduced by Isola et al. [36, 3, 87, 80] omitted noise input in its generator. Instead, it learned a direct mapping from an observed image y to an output image $G(y)$, such as translating grayscale images into colorized versions. As a follow-up, pix2pixHD [70] enhanced this framework by leveraging cGANs with feature matching loss to enable high-resolution image synthesis and semantic manipulation. The learning problem was reformulated as a multi-task problem with discriminators specialized in handling high-resolution details.

Other innovations include robust cGANs proposed by Chrysos et al. [12] and methods addressing noisy labels, as discussed by Thekumparampil et al. [66]. Conditional CycleGAN [43] added cyclic consistency constraints to cGANs, while Mode Seeking GANs (MSGANs) [45] introduced a regularization term to mitigate mode collapse.

Figure 4.

Multimodal image translation [97, 92, 86, 83, 85] enables mapping input images to a distribution of outputs in the target domain, retaining fidelity to the input. This paradigm addresses the mode collapse problem [22, 2, 24, 84], where generators produce repetitive outputs irrespective of input variations.

BicycleGAN [99] combined cVAE-GAN [27, 34, 35] and cLR-GAN [11, 14, 16] to achieve supervised multimodal translation. It systematically tackled mode collapse and generated diverse outputs. Similarly, PixelNN [5] employed pixelwise nearest-neighbor matching to condition translations on multiple exemplars, enabling controllable outputs. Another solution is disentangled representation learning [11, 26, 31, 13], which decomposes features into domain-invariant (content) and domain-specific (style) components. Gonzalez-Garcia et al. [21] extended this approach by dividing representations into shared and exclusive parts, enabling bi-directional multimodal translation and interpolation across domains. **Applications and Future Directions**. Conditional GANs and their derivatives have been applied to various tasks, including text-to-image synthesis [60, 88], panoramic image generation [20], and exemplar-based synthesis [97]. The ability to translate low-resolution to high-resolution images has enabled applications in super-resolution [78], benefiting autonomous driving, medical imaging, and more.

These advancements mark significant progress in image synthesis, with future directions likely focusing on improving generalization, computational efficiency, and application diversity. The integration of disentangled representation learning and multimodal capabilities will further enhance the adaptability of these systems across domains.

Advancements in Neural Radiance Fields (NeRF): Framework, Applications, and Innovations. Neural Radiance Fields (NeRF) have emerged as a transformative technology in the realm of 3D modeling, rendering, and dynamic scene representation. This section explores the foun-

Figure 5.

dational framework of NeRF, its core applications, and the latest advancements in its evolving landscape.

At its core, NeRF represents a 3D point and a 2D viewing direction, mapping them to an emitted color and volume density. This mapping is parameterized through various techniques, including Multilayer Perceptrons (MLPs) [48, 6], discrete voxel grids [17, 65], tensor decompositions [9], or hash mapping methods [50]. The rendering process computes the color of each pixel via volume rendering [15] along a ray originating from the camera. Volume rendering aggregates color and density along the ray to generate a 2D image projection, offering a powerful tool for novel view synthesis. The optimization process involves minimizing the reconstruction error between multi-view images and the synthesized outputs, effectively reconstructing the 3D scene geometry and appearance.

NeRF's primary application lies in 3D reconstruction, enabling high-quality novel view synthesis. Beyond static reconstructions, NeRF has been expanded to incorporate various editing capabilities. Object manipulation allows maneuvering and editing objects within a scene [74, 77]. Style transfer applies artistic styles to scenes and objects, enabling unique visual representations [91, 40]. Text-based editing provides the ability to modify visual content guided by natural language descriptions [25, 76]. Scene relighting alters illumination conditions in 3D environments, enhancing their aesthetic and contextual appearance [46]. NeRF's integration into generative models [10, 23, 9, 69, 49, 1, 90] facilitates the creation of 3D assets and serves as a distillation target for 2D foundation models. This integration enables the generation of 3D objects from textual descriptions [58, 30, 72], making NeRF invaluable for applications such as pose estimation, tracking, and navigation.

While early NeRF models focused on static scenes, subsequent innovations extended its utility to dynamic scenes [64, 56, 61, 94, 79, 95]. Articulated body movements are captured by models like Neural Body [57], which use vertex-based deformable human body models with latent codes for frame interpolation. NerFace [19] represents facial expressions dynamically by integrating per-frame latent codes with a 76-dimensional morphable model [67]. Non-rigid transformations are handled by Nerfies [54] and HyperNeRF [55], which extend canonical spaces to higher dimensions for robust dynamic scene representation. HDR imaging has also been advanced with RawNeRF [47], adapting NeRF for linear color spaces, and HDR-NeRF [28], synthesizing HDR images from LDR training data with variable exposure times. These dynamic adaptations extend NeRF's capabilities to motion capture, video rendering, and scene animation, significantly enhancing its versatility across industries.

NeRF has also been applied to fundamental image processing tasks, offering significant improvements in various domains. For denoising and deblurring, methods like DeblurNeRF [44] address motion blur, enhancing image clarity. Super-resolution techniques such as NeRF-SR [68] generate high-resolution views from low-quality inputs. Semantic label synthesis is enabled by models like Semantic-NeRF [96] and Fig-NeRF [73], which can generate semantic labels for novel views, advancing applications in robotics and autonomous systems. Additionally, the extraction of 3D surfaces and geometries from NeRF has seen notable developments. Mesh generation is facilitated by algorithms such as marching cubes [42], which produce detailed mesh representations. Density thresholding methods like UNISURF [53] refine baseline NeRF density thresholds for improved surface reconstruction. Techniques leveraging Signed Distance Functions (SDFs) [71, 4, 18] integrate scene geometries with volume rendering for precise surface modeling. Together, these advancements highlight NeRF's expanding role in advancing 3D modeling and image processing.

3. Methodology

The methodologies underpinning Neural Style Transfer (NST) and Image Synthesis represent the confluence of neural networks, mathematical optimization, and creative applications. This section delineates the techniques, architectures, and processes used to implement these advanced methods, providing a comprehensive overview of their respective frameworks.

3.1. Neural Style Transfer (NST)

Neural Style Transfer (NST) is designed to synthesize an image that combines the content of one image with the style of another. This process relies on extracting and recombining visual features through Convolutional Neural Networks (CNNs).

3.1.1 Core Components of NST

The implementation of NST involves three main images:

- **Content Image:** The image whose structural details (e.g., objects and layout) are to be preserved.

- **Style Image:** The image whose artistic elements (e.g., color, texture, brush strokes) are to be transferred.
- **Generated Image:** The synthesized image created through iterative optimization.

3.1.2 Feature Extraction

The process begins by passing the content, style, and generated images through a pre-trained CNN, typically VGG-19. Different layers of the network extract features that represent high-level abstractions for content and low-level patterns for style. Specifically:

- **Content Features:** Extracted from higher layers of the network, which capture semantic information about the input.
- **Style Features:** Extracted from lower layers, representing textures, patterns, and other stylistic attributes.

3.1.3 Loss Functions

The quality of the generated image is determined by minimizing a total loss function, comprising:

- **Content Loss:**

$$\mathcal{L}_{\text{content}} = \frac{1}{2}\sum_{i,j}\left(F_{ij}^{c} - F_{ij}^{g}\right)^2,$$

  where F_{ij}^{c} and F_{ij}^{g} denote the feature representations of the content and generated images, respectively, at a specific layer.

- **Style Loss:**

$$\mathcal{L}_{\text{style}} = \frac{1}{4N^2M^2}\sum_{i,j}\left(G_{ij}^{s} - G_{ij}^{g}\right)^2,$$

  where G_{ij}^{s} and G_{ij}^{g} represent the Gram matrices of the style and generated images, respectively.

- **Total Loss:**

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{content}} + \beta\mathcal{L}_{\text{style}},$$

  where α and β are weighting parameters that balance content preservation and style transfer.

3.1.4 Optimization

Gradient descent is used to iteratively adjust the pixel values of the generated image. Backpropagation minimizes the total loss, ensuring that the generated image increasingly resembles the style and content of the respective input images.

3.2. Image Synthesis via Generative Adversarial Networks (GANs)

Image synthesis involves generating novel images from scratch, often guided by noise or latent variables. This is achieved through Generative Adversarial Networks (GANs), comprising two neural networks: a generator and a discriminator.

3.2.1 GAN Architecture

- **Generator:** The generator $G(z)$ takes random noise z as input and generates synthetic images. Its architecture typically consists of:
  - Transposed convolutional layers for upsampling.
  - Batch normalization for stabilizing training.
  - Activation functions like ReLU or LeakyReLU for non-linearity.
- **Discriminator:** The discriminator $D(x)$ evaluates the authenticity of an input image, distinguishing between real and generated samples. It employs:
  - Convolutional layers for feature extraction.
  - Sigmoid activation for binary classification.

3.2.2 Adversarial Training Process

GANs employ a min-max optimization framework, where the generator and discriminator are trained in opposition:

- **Generator Objective:** Minimize the discriminator's ability to identify fake images:

$$\mathcal{L}_G = \mathbb{E}_z[\log(1 - D(G(z)))].$$

- **Discriminator Objective:** Maximize its ability to classify real and fake images:

$$\mathcal{L}_D = \mathbb{E}_x[\log D(x)] + \mathbb{E}_z[\log(1 - D(G(z)))].$$

The overall objective function is:

$$\min_G \max_D \mathcal{L}(G, D) = \mathbb{E}_x[\log D(x)] + \mathbb{E}_z[\log(1 - D(G(z)))].$$

3.2.3 Advanced GAN Techniques

To improve the quality and diversity of generated images, several advanced techniques are employed:

- **Conditional GANs (cGANs):** Incorporate additional information, such as class labels or text descriptions, to control image generation.

- **StyleGAN:** Introduces a style-based generator architecture, enabling fine-grained control over image attributes.

- **Progressive GANs:** Gradually increase image resolution during training for high-quality outputs.

3.3. Hybrid Approach: Combining NST and GANs

The intersection of Neural Style Transfer and Image Synthesis offers a powerful framework for creating stylized synthetic images:

- Generate a base image using GANs, capturing structural details and realism.

- Apply Neural Style Transfer to overlay artistic patterns and textures onto the generated image.

3.3.1 Applications of Hybrid Methods

This combined approach has practical applications in:

- **Artistic Content Creation:** Producing stylized artwork with unique and compelling aesthetics.

- **Enhanced Data Augmentation:** Diversifying synthetic datasets with stylistic variations.

- **Interactive Tools:** Developing applications for artists and designers to experiment with styles and concepts in real time.

3.4. Evaluation Metrics

The effectiveness of NST and GAN-based image synthesis is evaluated using:

- **Perceptual Quality:** Human evaluations to assess the aesthetic appeal and fidelity of generated images.

- **Fréchet Inception Distance (FID):** Measures the similarity between distributions of real and generated images.

- **Style Consistency:** Quantifies how closely the stylized image matches the reference style.

3.5. Conclusion of Methodology

The methodologies underlying Neural Style Transfer and Image Synthesis highlight the synergy between mathematical precision and creative innovation. By leveraging advanced neural networks, optimization techniques, and hybrid approaches, these methods pave the way for groundbreaking applications in art, design, and beyond. Their continued evolution promises to redefine the landscape of visual creativity and computational aesthetics.

References

[1] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *CVPR*, 2023. 6

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017. 5

[3] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. Compositional gan: Learning image-conditional binary composition. *International Journal of Computer Vision*, 128(10):2570–2585, 2020. 5

[4] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6290–6301, 2022. 6

[5] Aayush Bansal, Yaser Sheikh, and Deva Ramanan. Pixelnn: Example-based image synthesis. In *International Conference on Learning Representations*, 2018. 5

[6] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, 2021. 6

[7] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059, 2019. 2

[8] Jinming Cao, Oren Katzir, Peng Jiang, Dani Lischinski, Danny Cohen-Or, Changhe Tu, and Yangyan Li. Dida: Disentangled synthesis for domain adaptation, 2018. 2

[9] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 6

[10] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2021. 6

[11] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Neural Information Processing Systems*, pages 2172–2180, 2016. 5

[12] Grigorios G Chrysos, Jean Kossaifi, and Stefanos Zafeiriou. Robust conditional generative adversarial networks. *arXiv preprint arXiv:1805.08657*, 2018. 5

[13] Emily L Denton and vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4414–4423. Curran Associates, Inc., 2017. 5

[14] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 5

[15] Robert A Drebin, Loren Carpenter, and Pat Hanrahan. Volume rendering. *ACM Siggraph Computer Graphics*, 1988. 6

[16] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016. 5

[17] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 6

[18] Qiancheng Fu, Qingshan Xu, Yew-Soon Ong, and Wenbing Tao. Geo-neus: geometry-consistent neural implicit surfaces learning for multi-view reconstruction. In *Advances in Neural Information Processing Systems*, 2022. 6

[19] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 6

[20] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017. 5

[21] Abel Gonzalez-Garcia, Joost Van De Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *Advances in neural information processing systems*, pages 1287–1298, 2018. 5

[22] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks, 2017. 5

[23] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv:2110.08985*, 2021. 6

[24] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Neural Information Processing Systems*, pages 5767–5777, 2017. 5

[25] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv:2303.12789*, 2023. 6

[26] I. Higgins, Loïc Matthey, A. Pal, Christopher P. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 5

[27] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 5

[28] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. Hdr-nerf: High dynamic range neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18398–18408, 2022. 6

[29] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 4, 5

[30] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, 2022. 6

[31] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018. 5

[32] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865, 2017. 2

[33] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014. 2

[34] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:1, 2014. 5

[35] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016. 5

[36] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9455–9464, 2018. 5

[37] Jianxin Lin, Yingxue Pang, Yingce Xia, Zhibo Chen, and Jiebo Luo. Tuigan: Learning versatile image-to-image translation with two unpaired images. In *European Conference on Computer Vision*, pages 18–35. Springer, 2020. 2

[38] Jianxin Lin, Yingce Xia, Sen Liu, Tao Qin, and Zhibo Chen. Zstgan: An adversarial approach for unsupervised zero-shot image-to-image translation. *arXiv preprint arXiv:1906.00184*, 2019. 2

[39] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. In *Advances in neural information processing systems*, pages 2590–2599, 2018. 2

[40] Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric P Xing. Stylerf: Zero-shot 3d style transfer of neural radiance fields. In *CVPR*, 2023. 6

[41] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[42] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *ACM Siggraph Computer Graphics*, 1987. 6

[43] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Conditional cyclegan for attribute guided face image generation. *arXiv preprint arXiv:1705.09966*, 2017. 5

[44] Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V Sander. Deblur-nerf: Neural radiance

fields from blurry images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12861–12870, 2022. 6

[45] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1429–1437, 2019. 5

[46] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 6

[47] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16190–16199, 2022. 6

[48] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 6

[49] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *CVPR*, 2023. 6

[50] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ToG*, 2022. 6

[51] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[52] Aamir Mustafa and Rafał K. Mantiuk. Transformation consistency regularization – a semi-supervised paradigm for image-to-image translation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 599–615, Cham, 2020. Springer International Publishing. 2

[53] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 6

[54] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 6

[55] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), dec 2021. 6

[56] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14314–14323, October 2021. 6

[57] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 6

[58] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv:2209.14988*, 2022. 6

[59] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in neural information processing systems*, pages 3546–3554, 2015. 2

[60] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text-to-image synthesis. In *ICML*, 2016. 5

[61] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yan-Pei Cao, Tao Yu, and Yebin Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15872–15882, June 2022. 6

[62] Mingguang Shi and Bing Zhang. Semi-supervised learning improves gene expression-based prediction of cancer recurrence. *Bioinformatics*, 27(21):3017–3023, 2011. 2

[63] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 2

[64] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34:12278–12291, 2021. 6

[65] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, 2022. 6

[66] Kiran K Thekumparampil, Ashish Khetan, Zinan Lin, and Sewoong Oh. Robustness of conditional gans to noisy labels. In *Neural Information Processing Systems*, pages 10271–10282, 2018. 5

[67] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 6

[68] Chen Wang, Xian Wu, Yuan-Chen Guo, Song-Hai Zhang, Yu-Wing Tai, and Shi-Min Hu. Nerf-sr: High quality neural radiance fields using supersampling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6445–6454, 2022. 6

[69] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR*, 2023. 6

[70] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 5

[71] Yiqun Wang, Ivan Skorokhodov, and Peter Wonka. Hf-neus: Improved surface reconstruction using high-frequency details. In *Advances in Neural Information Processing Systems*, 2022. 6

[72] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv:2305.16213*, 2023. 6

[73] Christopher Xie, Keunhong Park, Ricardo Martin-Brualla, and Matthew Brown. Fig-nerf: Figure-ground neural radiance fields for 3d object category modelling. In *2021 International Conference on 3D Vision (3DV)*, pages 962–971. IEEE, 2021. 6

[74] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *ICCV*, 2021. 6

[75] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 2

[76] Lu Yu, Wei Xiang, and Kang Han. Edit-diffnerf: Editing 3d neural radiance fields using 2d diffusion model. *arXiv:2306.09551*, 2023. 6

[77] Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove. Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering. In *CVPR*, 2021. 6

[78] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 5

[79] Fangneng Zhan, Lingjie Liu, Adam Kortylewski, and Christian Theobalt. General neural gauge fields. In *The Eleventh International Conference on Learning Representations*, 2023. 6

[80] Fangneng Zhan and Shijian Lu. Esir: End-to-end scene text recognition via iterative image rectification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2059–2068, 2019. 5

[81] Fangneng Zhan, Shijian Lu, and Chuhui Xue. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–266, 2018. 2

[82] Fangneng Zhan, Chuhui Xue, and Shijian Lu. Ga-dan: Geometry-aware domain adaptation network for scene text detection and recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9105–9115, 2019. 2

[83] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 5

[84] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Kaiwen Cui, Changgong Zhang, and Shijian Lu. Auto-regressive image synthesis with integrated quantization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 110–127. Springer, 2022. 5

[85] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Lingjie Liu, Adam Kortylewski, Christian Theobalt, and Eric Xing. Multimodal image synthesis and editing: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 5

[86] Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Rongliang Wu, and Shijian Lu. Modulated contrast for versatile image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18280–18290, 2022. 5

[87] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3653–3662, 2019. 5

[88] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision*, pages 5907–5915, 2017. 5

[89] Jiahui Zhang, Fangneng Zhan, Christian Theobalt, and Shijian Lu. Regularized vector quantization for tokenized image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18467–18476, 2023. 2

[90] Jiahui Zhang, Fangneng Zhan, Rongliang Wu, Yingchen Yu, Wenqing Zhang, Bai Song, Xiaoqin Zhang, and Shijian Lu. Vmrf: View matching neural radiance fields. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6579–6587, 2022. 6

[91] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *ECCV*, 2022. 6

[92] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 5

[93] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *Advances in Neural Information Processing Systems*, pages 2365–2374, 2018. 2

[94] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7753, 2022. 6

[95] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for

human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15893–15903, 2022. 6

[96] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 6

[97] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11465–11475, 2021. 5

[98] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*, pages 2223–2232, 2017. 2

[99] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Neural Information Processing Systems*, pages 465–476, 2017. 5