Fine-Grained Guidance for Image Generation

Hidetomo Sakaino
FSOFT-FJP-FAI-ABC
Mita, Minato-ku, Tokyo
HidetomoS@fpt.com

Nam Nguyen
FSOFT-FJP-FAI-ABC
Mita, Minato-ku, Tokyo
namnx21@fpt.com

Abstract

Fine-grained image analysis (FGIA) plays a crucial role in precise object understanding, enabling detailed segmentation and categorization of specific object components, e.g., valve elements or automotive parts. Traditional FGIA methods, which primarily rely on geometric descriptors, often struggle with variable real-world conditions, i.e., illumination changes, weathering effects, and diverse viewing angles. To address these limitations, we proposed a multi-modal framework that combines generative AI with fine-grained visual guidance, jointly performing part-level alignment and detailed image synthesis. Our proposed method leverages advanced computer vision techniques for precise subclass recognition and segmentation while integrating generative models capable of generating accurate multi-view images from single-view inputs. This synthesis process incorporates explicit functional constraints and clear semantic alignments, significantly improving segmentation accuracy by mitigating mismatches between visual features and semantic cues. Experimental results confirm that our approach notably outperforms state-of-the-art methods in both fine-grained segmentation accuracy and robust image generation under challenging conditions.

1. Introduction

Fine-grained image analysis (FGIA) represents a critical challenge in computer vision (CV) and pattern recognition, significantly impacting real-world applications that demand precise differentiation among closely related object categories, such as specific valve types or car models [21, 46, 47]. The inherent complexity of FGIA arises primarily from subtle inter-class variations combined with pronounced intra-class differences, necessitating accurate localization and extraction of discriminative features from extensive image datasets [45–47]. Recent advancements in FGIA have extended to include refined subclass identification, leveraging visual references to improve differentiation among highly similar categories [7, 17, 18, 22].

Traditional and state-of-the-art (SOTA) FGIA techniques, such as Local Binary Patterns (LBP), Histogram of Gradients (HOG), and SIFT [2], along with deep learning-based approaches, typically depend on geometric descriptors (e.g., edges, salient keypoints). These methods seek robustness to occlusions and limited viewing angles (< 45 degrees), focusing predominantly on object recognition. Nonetheless, practical conditions frequently introduce significant challenges due to variable lighting, occlusions, surface degradation, and diverse viewpoints, complicating reliable subclass identification. Notably, standard reference data predominantly features pristine objects, contrasting sharply with real-world images subject to environmental deterioration.

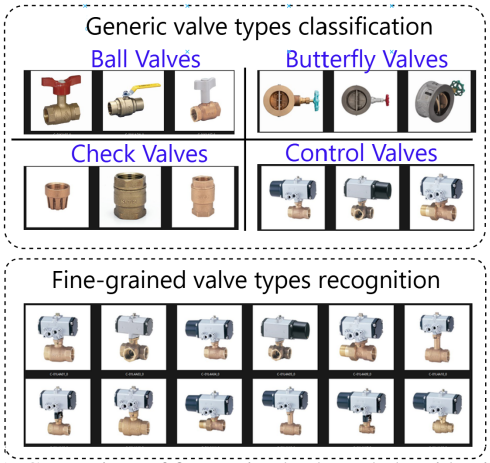

Figure 1. Comparison of fine-grained valve subclass identification against generic classification.

Most existing FGIA methodologies employ a purely 2D-to-2D approach [57], prevalent in both academic research and industry, e.g., Google Vertex. Meanwhile, the integration of 2D-to-3D generative models to bolster fine-grained identification remains relatively underexplored. Current methods struggle under substantial appearance variations arising from environmental or viewpoint shifts, limiting their capability to consistently achieve fine-grained subclass recognition. Typically, generative AI (Gen-AI) and

FGIA methodologies have evolved independently, lacking integrated frameworks capable of simultaneously addressing subclass differentiation and robust visual representation.

Addressing these critical limitations, this paper proposes a novel FGIA framework that synergistically integrates spatio-temporal analysis, conventional CV techniques, and Gen-AI for robust and precise fine-grained subclass identification. Our methodology combines advanced CV-based recognition [30, 31] and classification techniques [20, 39], enhancing segmentation and differentiation between reference and target subclasses. Furthermore, the framework employs Gen-AI models, such as Stable Diffusion [26, 42, 44, 48], to generate detailed 3D representations from 2D images, enabling robust multi-view analysis across varying conditions, including occlusion, corrosion, and viewpoint changes.

The primary contributions of this research are fivefold:

- First, a comprehensive FGIA framework integrating spatio-temporal analysis with Gen-AI to overcome limitations of current static 2D approaches, enabling detailed subclass-specific image generation.
- Second, novel constraints designed to address inconsistencies between visual and textual subclass representations, thereby significantly enhancing subclass identification accuracy.
- Third, utilization of temporal image sequences to mitigate issues arising from environmental degradation, significantly improving the robustness and reliability of fine-grained recognition.
- Fourth, a novel multi-modal vision-language approach addressing previous gaps in semantic understanding by generating comprehensive and precise subclass descriptions.
- Finally, empirical validation demonstrating superior subclass identification performance of our proposed method on realistic datasets, highlighting substantial improvements achieved through 2D-to-3D generative modeling and functional segmentation techniques compared to current state-of-the-art FGIA methodologies.

2. Related work

This section describes recent advancements in generative modeling for computer vision, emphasizing the integration of vision-language models (VLMs) and large language models (LLMs), particularly for text-to-image generation. VLMs utilize extensive multimodal datasets to capture complex image-text correlations, while LLMs contribute contextual understanding, significantly enhancing generated image coherence and relevance. We critically analyze significant contributions from previous works, highlighting the strengths and limitations of existing approaches.

2.1. Generative Vision-Language Models

Recent surveys have summarized core architectures, training methods, and applications of VLMs and generative models, providing valuable overviews of model evolution from traditional vision systems to modern web-scale architectures. These works categorize models comprehensively, detailing their theoretical foundations and practical applications across various vision tasks, particularly in text-to-image synthesis and image editing [12, 35, 54]. Key foundational models include transformer-based systems such as DALL-E [33] and CLIP [32], as well as latent diffusion models like Stable Diffusion [36]. These models significantly advanced image realism, synthesis quality, and computational efficiency [8, 15, 41].

However, existing literature primarily offers theoretical insights, inadequately addressing practical challenges like dataset realism, diversity, and domain adaptation. A notable performance gap remains between synthetic and real-world data, with limitations in capturing complex scene dynamics and generalizing effectively across diverse real-world conditions. Consequently, new methodologies are required to improve synthetic image realism, diversity, and generalization capabilities for practical deployment.

2.2. Vision-Guided and Language-Guided Methods

Contemporary methods combining VLMs and LLMs have advanced image generation realism, robustness, and personalization significantly [4, 5, 9, 28, 34, 37, 50, 52, 53]. Models featuring modality collaboration and adaptive modules have set new benchmarks in image synthesis. Additionally, language-guided generative frameworks have enhanced controllability, precision, and contextual relevance, leveraging textual instructions for improved image editing and synthesis control [3, 6, 23, 43, 49, 55].

Recent progress in generative modeling has revealed a growing emphasis on adaptability and task-specific capabilities. Advances such as plausibility-aware 3D mesh deformation, personalized generation, open-vocabulary segmentation, and robust detection of synthetic content reflect this shift [11, 19, 24, 27, 29, 38, 51]. However, critical challenges remain unresolved. Current models struggle to synthesize images that faithfully capture physical phenomena—such as material corrosion, realistic shadow casting, and consistent physical scaling—hindering their application in domains where physical realism is essential. Additionally, issues related to computational efficiency, ambiguity in textual conditioning, scalability, and ethical considerations continue to limit deployment. Addressing these gaps demands new approaches that integrate physical priors, improve generative accuracy under ambiguous supervision, and enforce ethical and scalable design—paving the way for robust and practically viable generative systems in vision tasks.

3. Proposed method

This section describes in detail the proposed method as shown in Figure 2. Given an input image x of an industrial object, we first obtain its latent–noise representation by running the *forward* phase of a deterministic DDIM scheduler [40] for T steps, yielding ε_T. Image captions c enumerate fine-grained attributes such as material, color, and component geometry, respectively.

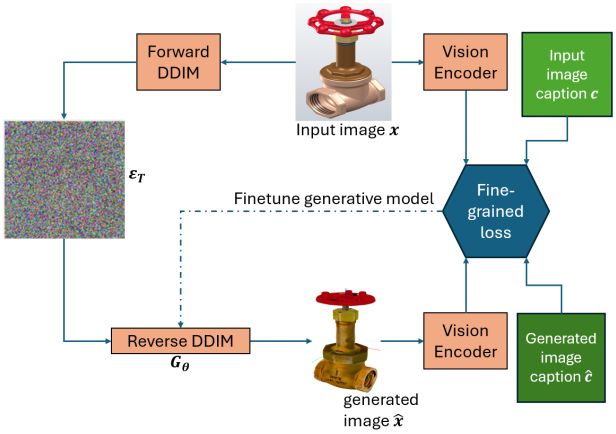

Figure 2. Proposed method for image generation

Starting from ε_T, a diffusion generator G_θ [16] then executes the *reverse* DDIM trajectory to synthesize an image $\hat{x} = G_\theta(\varepsilon_T)$, which is immediately re-captioned by C to obtain $\hat{c}$.

The four signals $(x, \hat{x}, c, \hat{c})$ are coupled through a **fine-grained loss** $\mathcal{L}_{\text{FG}}$. The resulting gradient is back-propagated *only through* the generator parameters θ; the captioner C and the forward DDIM path remain frozen. Because the same deterministic scheduler drives both the forward and reverse processes, each training step forms an exact cycle, eliminating stochastic mismatch and allowing the model to concentrate capacity on attribute-level fidelity—as required, for example, to distinguish a rust-flecked bronze valve from a clean brass one.

Let f_θ be the vision encoder that maps an image to an embedding $z = f_\theta(x)$, and let g_ϕ be a text encoder that maps the caption to $t = g_\phi(c)$. For every original image, x we synthesize two additional variants: x' that mimics *temporal* deterioration (e.g. rust, paint fading) and x'' that mimics *environmental* variation (e.g. glare or low light). The overall training objective is a weighted sum of four terms

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ctr}}\mathcal{L}_{\text{contrast}} + \lambda_{\text{det}}\mathcal{L}_{\text{detail}} + \lambda_{\text{tmp}}\mathcal{L}_{\text{temporal}} + \lambda_{\text{env}}\mathcal{L}_{\text{env}}, \tag{1}$$

where the λ coefficients balance the influence of each component.

(i) Contrastive Loss $\mathcal{L}_{\text{contrast}}$. We adopt the margin-based formulation of Hadsell *et al.* [13]:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{2} y D^2 + \frac{1}{2}(1-y)\big[\max(0, m - D)\big]^2, \tag{2}$$

where $D = \|z_i - z_j\|_2$, $y = 1$ if the pair (x_i, x_j) belongs to the same fine-grained subclass and 0 otherwise, and m is a fixed margin.

(ii) Detail-Alignment Loss $\mathcal{L}_{\text{detail}}$. To force individual image *patches* to align with the corresponding textual *tokens*, we split x into K patches $\{p_k\}$ with embeddings $\{z_k\}$ and obtain token embeddings $\{t_k\}$. Using an all-pairs InfoNCE over the similarity matrix $S_{kj} = \langle z_k, t_j \rangle$, we define

$$\mathcal{L}_{\text{detail}} = -\frac{1}{K}\sum_{k=1}^{K} \log \frac{\exp\big(S_{k,k}/\tau\big)}{\sum_{j=1}^{K} \exp\big(S_{k,j}/\tau\big)}, \tag{3}$$

where τ is a temperature hyper-parameter. Patch embeddings are extracted from a ViT backbone [10] or its hierarchical variant Swin-T [25].

(iii) Temporal Consistency Loss $\mathcal{L}_{\text{temporal}}$. For the pair (x, x') that differs only by time-based degradation, we impose

$$\mathcal{L}_{\text{temporal}} = \big\| f_\theta(x) - f_\theta(x') \big\|_2^2. \tag{4}$$

If ground-truth masks identifying the corroded region are available, the loss can be evaluated only on those pixels; we also experiment with an optional CycleGAN-style consistency term [56].

(iv) Environment-Invariant Loss $\mathcal{L}_{\text{env}}$. Let $A_e(\cdot)$ denote a photometric augmentation that simulates a specific environmental condition e (overexposure, glare, or low light). We minimize

$$\mathcal{L}_{\text{env}} = \mathbb{E}_e \, \big\| f_\theta(x) - f_\theta\big(A_e(x)\big) \big\|_2^2, \tag{5}$$

and optionally treat $A_e(x)$ as a positive sample in the contrastive term.

Weight Selection. We follow the guideline $\lambda_{\text{ctr}} = 1$, $\lambda_{\text{det}} \in [1, 2]$, $\lambda_{\text{tmp}} \in [0.2, 0.5]$, and $\lambda_{\text{env}} \in [0.2, 0.5]$, and employ GradNorm to adaptively rescale the λ values during training.

The composite objective (1) therefore enforces *instance-level discrimination*, *part-level correspondence*, *temporal stability*, and *environmental robustness* simultaneously, enabling the generator to reproduce fine-grained visual details with a level of fidelity previously unattainable by diffusion models.

4. Experiments

This section presents quantitative results for subtype identification—with and without our proposed data-augmentation method—alongside qualitative visualizations.

4.1. Subtype Identification Experiment

We assembled 872 images of six valve types—ball, check, butterfly, control, gate, and plug—with subtype counts ranging from 5 to 42 images each to evaluate fine-grained classification without augmentation. Augmenting the dataset by 50 % markedly enriched appearance and pose diversity, boosting model robustness and generalization.

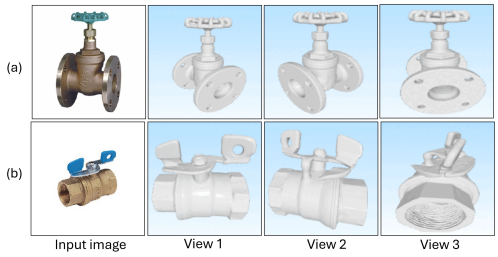

Figure 3. Generated images from different views

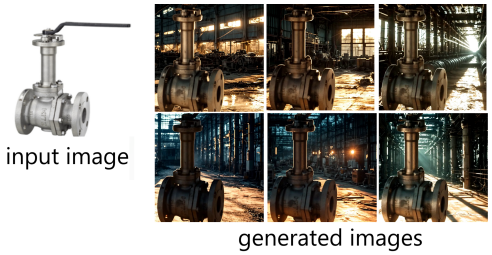

Figure 4. Generated images from different conditions

Table 1 shows classification accuracy (%) and per-image runtime for ResNet-101 [14] and FineLIP [1], with and without our augmentation. Augmentation consistently boosts accuracy by 9% —ResNet-101 from 76.12 to 85.3 and FineLIP from 79.65 to 88.7—while adding only 20 ms per image in the ResNet pipeline. FineLIP's richer augmentation achieves the highest accuracy but a 2.5 s per-image cost, underscoring the expressiveness–runtime trade-off.

Table 1. Accuracy Comparison of ResNet-101 [14], FineLIP [1] with and without data augmentation using proposed method

Method	Without(%)	With(%)	Cost (ms)
ResNet-101 [14]	76.12	85.3	20
FineLIP [1]	79.65	88.7	2500

The improved performance of the proposed method can be attributed to the advanced feature integration and dynamic modulation strategies that better leverage multi-view information and time-varying conditions. This enhanced approach effectively bridges the gap between subtle visual variations and robust feature extraction, thereby providing a significant performance boost over existing methods in challenging fine-grained classification tasks.

4.2. Experiment on detailed image generation

This subsection describes a qualitative evaluation of our model's ability to generate highly detailed images of industrial valve components from fine-grained textual prompts. **Prompt:** *A T-shaped valve body with raised-face flanges reveals a 2-inch, Class-300 globe (or gate) valve, its stem and bonnet just peeking into view. Crisp cast-in markings—KITZ, 2, STEEL, 300, WCB—are joined by a small boxed code and "C"/"1" identifiers on the bosses. The unpainted, light-grey carbon-steel surface bears a shot-blasted matte texture and machining marks on the flange faces. Eight hex-head studs and nuts clamp the bonnet, their edges darkening with tarnish and early oxidation. Cinnamon-brown rust halos around bolt holes, faint streaks beneath lettering, and reddish pits on the lower flange rim signal early-stage corrosion, while the rest of the body remains metallic grey.* Figure 5 shows that the generated images capture the prompt's specified details; however, some elements remain unrealistic—for example, the characters embossed on the valve body are not rendered convincingly.

Figure 5. Generated images: (a) the first inference, (b) the second inference

5. Conclusion

This paper has presented a powerful generative data-augmentation strategy for fine-grained image analysis that injects photorealistic, semantically rich variants directly into training to improve subclass recognition and detailed image generation. Our proposed method has posed conditions at a diffusion model with fine-grained textual descriptors—capturing illumination shifts and time-varying —to synthesize realistic multi-view samples from single-view images, then aligns them with real data in a unified visual–semantic embedding. This tight coupling of generative synthesis and semantic guidance yields far greater appearance and pose diversity than standard augmentation schemes, driving a substantial leap in robustness without any bespoke post-processing. Future work will incorporate physical scale estimation into this framework.

References

[1] Mothilal Asokan, Kebin Wu, and Fatima Albreiki. Finelip: Extending clip's reach via fine-grained alignment with longer text inputs, 2025. 4

[2] Simon Bilik and Karel Horak. Sift and surf based feature extraction for the anomaly detection, 2022. 1

[3] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. 2

[4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Chen, and William T. Freeman Shi. Muse: Text-to-image generation via masked generative transformers, 2023. 2

[5] Aniruddha Chatterjee, Shalini Gupta, and Yash Patel. Robustness of generative models using language guidance for low-level vision tasks. In *CVPR Workshop on Generative Models for Computer Vision*, 2024. 2

[6] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based probing of frozen vision-language models, 2023. 2

[7] Antonio Criminisi, Ian Reid, and Andrew Zisserman. Single view metrology. *International Journal of Computer Vision*, 40(2):123–148, 2000. 1

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8780–8794, 2021. 2

[9] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhewei Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers, 2021. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[11] Yang Du, Yuheng Li, and Haotian Zhang. Intrinsic lora: A generalist approach for discovering knowledge in generative models. In *CVPR Workshop on Generative Models for Computer Vision*, 2024. 2

[12] Arindam Ghosh. Exploring the frontier of vision-language models: A survey of current methodologies and future directions, 2024. 2

[13] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1735–1742, 2006. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 4

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[17] Martin Hofmann, Marco Seeland, and Patrick Mäder. Efficiently annotating object images with absolute size information using mobile devices. *International Journal of Computer Vision*, 127(2):207–224, 2019. 1

[18] Abhishek Kar, Shubham Tulsiani, João Carreira, and Jitendra Malik. Amodal completion and size constancy in natural scenes. *CoRR*, abs/1509.08147, 2015. 1

[19] Laurynas Karazija, Haotian Zhang, and Yuheng Li. Diffusion models for open-vocabulary segmentation. In *CVPR Workshop on Generative Models for Computer Vision*, 2024. 2

[20] Jashanpreet Kaur and Gurpreet Singh. Ai meets astronomy: Efficientb0- powered classification of ai-synthesized celestial objects using spacenet. In *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, pages 870–875, 2025. 2

[21] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. In *Second Workshop on Fine-Grained Visual Categorization (FGVC2), in conjunction with CVPR*, Portland, OR, 2013. 1

[22] Byeong-Uk Lee, Jianming Zhang, Yannick Hold-Geoffroy, and In So Kweon. Single view scene scale estimation using scale field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21435–21444, 2023. 1

[23] Yuheng Li, Haotian Zhang, Tianle Xu, Yiping Yang, Yihao Zhang, Jianzhu Zhu, and Ming-Hsuan Yang. Language-guided image generation with clip, 2023. 2

[24] Chen Liu, Haotian Zhang, and Yuheng Li. Turns: A new benchmark for robust detection of ai-generated videos. In *CVPR Workshop on Generative Models for Computer Vision*, 2024. 2

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 3

[26] Mariam Alaa Metwally and Milad Ghantous. Detecting generative ai in images. In *2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 214–220, 2024. 2

[27] Jiwan Park, Sanghyeon Kim, and Joonho Lee. Cat: Contrastive adapter training for personalized image generation. In *CVPR Workshop on Generative Models for Computer Vision*, 2024. 2

[28] Jiwan Park, Sanghyeon Kim, and Joonho Lee. Learning compositional language-based object detection with diffusion-based synthetic data. In *CVPR Workshop on Generative Models for Computer Vision*, 2024. 2

[29] Hieu Pham, Haotian Zhang, and Yuheng Li. Robust concept erasure using task vectors. In *CVPR Workshop on Generative Models for Computer Vision*, 2024. 2

[30] Daniel Phillips and Patrick Hosein. On the detection of manipulated identification documents. In *2024 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, pages 1–6, 2024. 2

[31] Chuyue Qi, Zonglin Yang, and Yuxin Wen. Improved resnet-50 model for ai image recognition based on multi-scale attention mechanism. In *2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE)*, pages 825–829, 2024. 2

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision, 2021. 2

[33] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 2

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2

[35] Anushka Raut and Amritpal Singh. Generative ai in vision: A survey on models, metrics and applications, 2024. 2

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2

[38] Andrew Shih, Daniel Goldman, and Jonathan T. Barron. Extranerf: Visibility-aware view extrapolation of neural radiance fields with diffusion models. In *CVPR Workshop on Generative Models for Computer Vision*, 2024. 2

[39] Gurpreet Singh, Kalpna Guleria, and Shagun Sharma. A pretrained efficientnetv2b0 model for the accurate classification of fake and real images. In *2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1082–1086, 2024. 2

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[41] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2020. 2

[42] Phi-Ho Truong, Tien-Dung Nguyen, Xuan-Hung Truong, Nhat-Hai Nguyen, and Duy-Trung Pham. Employing a cnn detector to identify ai-generated images and against attacks on ai systems. In *2024 1st International Conference On Cryptography And Information Security (VCRIS)*, pages 1–6, 2024. 2

[43] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image generation, 2023. 2

[44] Muthaiah U, A. Divya, T.N. Swarnalaxmi, and Vidhyasagar BS. A comparative review of ai-generated vs real images and classification techniques. In *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, pages 141–147, 2024. 2

[45] Xiu-Shen Wei, Chen-Wei Xie, and Jianxin Wu. Mask-cnn: Localizing parts and selecting descriptors for fine-grained image recognition. *CoRR*, abs/1605.06878, 2016. 1

[46] Xiu-Shen Wei, Jianxin Wu, and Quan Cui. Deep learning for fine-grained image analysis: A survey. *CoRR*, abs/1907.03069, 2019. 1

[47] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8927–8948, 2022. 1

[48] Stuart Weir, Muhammad Shahbaz Khan, Naghmeh Moradpoor, and Jawad Ahmad. Enhancing ai-generated image detection with a novel approach and comparative analysis. In *2024 17th International Conference on Security of Information and Networks (SIN)*, pages 1–7, 2024. 2

[49] Hu Xu, Haotian Zhang, Yuheng Li, Yiping Yang, Yihao Zhang, Jianzhu Zhu, and Ming-Hsuan Yang. Versatile diffusion: Text, images and variations, all in one model, 2023. 2

[50] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yizhuo Shi, et al. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13947–13957, 2024. 2

[51] Sangwon Yoo, Jiwan Lee, and Sanghyeon Kim. Asplausible-as-possible: Plausibility-aware mesh deformation using 2d diffusion priors. In *CVPR Workshop on Generative Models for Computer Vision*, 2024. 2

[52] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, et al. Vector-quantized image modeling with improved vqgan, 2021. 2

[53] Jiahui Yu, Jing Yu Koh, Vijay Vasudevan, Zirui Wang, Yuanzhong Xu, Gunjan Baid, Yinfei Yang, Alexander Ku, Yang Yang, Hongxia Yang, et al. Scaling autoregressive models for content-rich image completion, 2022. 2

[54] Huiping Zhang, Hongwei Liu, Rongrong Li, Yibing Liu, Chunyang Zhou, and Baocai Zhang. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6728–6748, 2024. 2

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 3

[57] Karel Zimmermann. Similarity among the 2d-shapes and the analysis of dissimilarity scores, 2022. 1