Artificial Intelligence in Creative Industries: Advances Prior to 2025

Nantheera Anantrasirichai, Fan Zhang, and David Bull
MyWorld
University of Bristol

Abstract

The rapid advancements in artificial intelligence (AI), particularly in generative AI and large language models (LLMs), have profoundly impacted the creative industries, enabling more innovative content creation, enhancing workflows, and democratizing access to creative tools. This paper explores these technological shifts, with particular focus on how those that have emerged since our previous review in 2022 have expanded creative opportunities and improved efficiency. These technological advancements have enhanced the capabilities of text-to-image, text-to-video, and multimodal generation technologies. In particular, key breakthroughs in LLMs have established new benchmarks in conversational AI, while advancements in image generators have revolutionized content creation. We also discuss the integration of AI into post-production workflows, which has significantly accelerated and improved traditional processes. Once content has been created, it must be delivered to its audiences the media industry is facing the demands of increased communication traffic due to creative content. We therefore include a discussion of how AI is beginning to transform the way we represent and compress media content. We highlight the trend toward unified AI frameworks capable of addressing and integrating multiple creative tasks, and we underscore the importance of human insight to drive the creative process and oversight to mitigate AI-generated inaccuracies. Finally, we explore AI's future potential in the creative sector, stressing the need to navigate emerging challenges and to maximize its benefits while addressing the associated risks.

Acknowledgements

This work has been funded by the UKRI MyWorld Strength in Places Programme (SIPF00006/1).

Contents

1 Introduction

The influence of artificial intelligence (AI) has grown dramatically over the past few years, particularly due to the rise of generative AI and large language models (LLMs). These advancements are widely regarded as beneficial by many countries, creating significant opportunities for growth (e.g. as outlined in the UK, by the Authority of the House of Lords [1]). These advances have also had significant direct and indirect impacts on the creative industries, influencing the direction of their growth. Generative AI, for instance, primarily focuses on generating new data that is not identical to the training data yet shares similarities with it. However, the cardinality of the training data can be huge, larger than what any individual human has ever encountered. The resulting output may therefore act as a new source of insipration.

AI tools also provide opportunities for a wider range of users to work more efficiently and effectively, with even greater creativity. Moreover, these new technologies not only influence creators, but they also enable new ways for audiences to experience art and culture [2].

A major breakthrough in generative AI has been led by OpenAI[1], an AI research and deployment company, with their introduction of Generative Pre-trained Transformer (GPT) models for LLMs. LLMs are specifically designed to understand and generate human language. They are characterized by their vast size in terms of parameters and the amount of training data used to create them. This breakthrough was particularly impactful when the company released ChatGPT in 2022, which was fine-tuned from a model in the GPT-3.5 series. ChatGPT is a conversational model that includes advanced safety features that mitigate the generation of inappropriate content. Several other LLM platforms were also developed contemporaneously, such as LaMDA and PaLM by Google AI, Ernie Bot by Baidu, and BLOOM by BigScience. Additionally, Anthropic launched Claude, the LLM trained specifically to be harmless and honest, leveraging reinforcement learning from human feedback (RLHF) - a technique used to train AI systems to appear more human [3]. Nonetheless, ChatGPT stands out as the most renowned, thanks to its quick and efficient responses, and notably its public accessibility, being available for free.

Another breakthrough in 2022 was in the area of text-to-image models. OpenAI achieved a significant milestone with DALL·E 2, producing impressive artworks and photorealistic images despite its limited language understanding. Midjourney by Midjourney, Inc., another well-known text-to-image generator, supports higher resolution images, up to 4096×4096 pixels. Stable Diffusion by Stability AI, for which the code and model weights are publicly available[2], allows developers and artists to further adapt AI to suit their own specific applications.

The next breakthrough happened in 2023 when OpenAI unveiled GPT-4, a significantly larger model with estimated 1.8 trillion parameters and improved performance compared to its predecessors [4]. However, this still represents less than 1% of the human brain's approximately 600 trillion synaptic connections[3]. GPT-4 is a multimodal large language model that can generate responses to both text and images. It incorporates DALL·E 3, enabling it to comprehend a much broader range of nuances and details than earlier versions. In March 2024, Claude 3 Opus by Anthropic was released, boasting multimodal capabilities in generating images, tables, graphs and diagrams. Moreover, Anthropic claims that Claude 3 Opus outperforms GPT-4 in generating human-like dialog and contextually aware responses. These rapid advances have, in

[1] https://openai.com/
[2] https://github.com/Stability-AI/stablediffusion
[3] https://www.rsb.org.uk/biologist-features/ai-versus-the-brain

turn, led the creative industries to face significant challenges. For example, DMG Media, the Financial Times, and Guardian Media Group have highlighted concerns about the potential impact on print journalism, particularly if AI tools reduce the need for users to click through to news websites, affecting advertising and subscription revenues [1]. There is also concern about 'AI-generated slop'—low-quality, mass-produced content created by AI that often lacks coherence or originality[4]. It is typically used for spam, search engines, or clickbait, and is criticized for cluttering the internet and undermining genuine human-created content.

The generation of videos is significantly more challenging for AI than generating images. In February 2024, Google announced Gemini 1.5 which had the capability to process approximately 8 times more data than GPT-4, opening opportunities for video and audio processing [5]. In the same month, OpenAI provided its first preview of Sora, a model capable of generating impressive realistic videos up to 1 minute long. Based on the videos released by OpenAI, Sora appears to outperform other text-to-video models. Sora is currently available to ChatGPT subscribers. A month later, Gemini 1.5 announced its support for native audio understanding in 180+ countries. With the emergence of these tools, together with the prospect of further advances, it is clear that video content creation will be a major beneficiary. This will further open up the media landscape for creativity and provide more opportunities for diverse storytellers, while also reducing production time. A recent example is the AI-generated Christmas commercial by Coca-Cola[6]. Such advertisements overcome the limitations of current technologies by using very short videos with rapid scene transitions, ensuring that any artifacts, such as unnatural fingers, are less apparent.

For the case of post-production workflows, generative AI may not have a direct impact, but the neural networks originally proposed for generative AI have been widely adapted to serve this purpose. This has led to significant improvements in both output quality and computational speed. Moreover, there is a noticeable trend towards adopting a unified framework rather than addressing individual tasks, as it better reflects real-world scenarios. For instance, natural history filmmaking involves challenging acquisition environments and high production standards. Filming often takes place in low light conditions, in the presence of heat haze, underwater or in adverse weather conditions. This often results in increased noise levels, focus issues, low contrast, color balance problems, and blurriness in the footage. In such cases, Unified models can offer advantages in generalizing to diverse tasks and providing flexibility. Take Painter by BAAI Vision [5] as an example, which employs an image pair as a task prompt (similar to a text prompt in LLMs), their model transfers the input image to produce a similar output as the task prompt, enabling it to undertake various tasks such as segmentation, low-light enhancement or rain removal.

While generative AI can facilitate and accelerate the creation and post-processing of digital media, there is an equivalent need to transmit or stream it efficiently to users. Although AI-based solutions have been proposed both for the enhancement of conventional video coding tools and for new compression frameworks, they are yet to be adopted in practical applications due to hardware constraints, complexity issues and a lack of standardization. Despite this, the latest learning-based video codecs have already demonstrated their potential to compete with conventional standardized video codecs and are being actively investigated in various standards bodies such as MPEG and AOM.

[4]https://reutersinstitute.politics.ox.ac.uk/news/ai-generated-slop-quietly-conquering-internet-it-threat-journalism-or-problem-
[5]https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/
[6]https://www.youtube.com/watch?v=4RSTupbfGog

Furthermore, in recent years, AI has also impacted our ability to assess and monitor the perceptual quality of visual media. Advances have included new model architectures based on different attention mechanisms and the application of LLMs, which evidently improve model generalization. New training methodologies have also been proposed based on weakly/unsupervised learning, which address issues associated with the limited availability of labeled training content.

One of the exciting aspects of using LLMs in the creative sector is that 'The human in the loop' [6] is simplified through text prompts, with sophisticated, multilingual language capabilities enabling artists to convey complex emotions and narratives. This is important because generative AI does produce mistakes, known as hallucinations. Human oversight is thus essential to correct this through reinforcement learning with feedback [7].

In this paper, the objective is to reveal to the reader, the latest technology advancements that have emerged since our previous review paper on AI in the creative industries (published in 2022) [8]. Compared to this earlier paper, which was written when most AI technologies were used as support tools, this updated review describes thesignificant disruptive shifts that have emerged over the past 3-4 years driven by generative AI and other recent AI-based technologies. Similar to [8], we first provide a high-level overview of current advanced AI technologies (Section 2), followed by a selection of key creative domain applications (Section 3) where current AI technologies are changing creative practice. Finally, we discuss the challenges and the future potential of AI associated with the creative industries (Section 4).

2 Current Advanced AI Technologies

This paper provides a review of AI in the creative industries, building on our previous publication in 2022 [8]. The reader is referred to that work for an introduction to AI, basic neurons, convolutional neural networks (CNNs), generative adversarial networks (GANs), recurrent neural networks (RNNs) and deep reinforcement learning (DRL). In this paper, we emphasize four key technologies that have grown in importance since 2022 that have had a significant impact on the creative industries. These are Transformers, Large language models (LLMs), Diffusion Models (DMs), and Implicit Neural Representations (INRs). It is important to note that, while these newer technologies are gaining prominence, those from previous generations remain in widespread use, often in conjunction with the newer ones. For instance, CNNs complement transformers since CNNs effectively capture local features and semantic meaning, while the attention mechanism in transformers capture global dependencies.

One important class within AI that has become dominant since our previous review comprises Foundation models (FMs). These were described by The Stanford Institute for Human-Centered Artificial Intelligence in 2021 [9] as "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks". Foundation models have been enabled by rapid advances in AI-oriented computing power and have been underpinned the emergence and success of Large Language Models, particularly following the launch of ChatGPT by OpenAI in 2022. ChatGPT has become the fastest-growing consumer software application in history[7].

These technologies are expanded on below.

[7]https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

2.1 Transformers

In 2017, Google AI introduced the concept of 'Transformer' architectures in their publication 'Attention Is All You Need' [10]. This work has since, been instrumental in the development and success of large language models alongside many other applications, including vision understanding [11], and multiple modality learning (e.g., Gato [12]).

Before the advent of transformers, natural language processing (NLP) was performed using recurrent neural networks (RNNs), processing data sequences sequentially. In contrast, the ability of transformers to capture long-range dependencies through self-attention mechanisms that extend across all words in the sequence, meant that the importance of different words could be established globally, understanding relationships regardless of their positions. This context-aware representation enables parallel processing of the entire sequence, making the transformers computationally efficient. A set of several attention layers running in parallel is called Multi-Head Attention.

The Transformer architecture, shown in Fig. 1 (a), comprises Encoder and Decoder sections, similar to many CNN-based generators. However the encoder is now a stack of identical layers, concatenating a multi-head self-attention mechanism and a fully connected feed-forward network. The decoder is also a stack of identical layers, in which each layer has additional sub-layer to perform multi-head attention over the output of the encoder stack.

Mathematically, the attention function is computed from inputs: query Q, keys K, and values V. The matrix of outputs of attention function is

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{1}$$

where d_k is a dimension of K. The term QK^T is Dot-Product Attention, which yields a high similarity value when the two words are closely related. If Q and K are from the same sentence, Eq. 1 refers to self-attention, but if Q and K are from different sentences, it is referred to as cross-attention. Within the network, multi-head attention is actually employed to concurrently process attention and enable the model to collectively focus on information from distinct representation subspaces at various positions through the learnable parameters Ws.

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O, \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \end{aligned} \tag{2}$$

It should be noted that attention modules are not solely used in transformers, but have also been successfully integrated into other deep learning architectures such as CNNs, used for image classification [13], object detection [14], and other computer vision tasks [15].

In 2020, the first successful training of a transformer encoder for image recognition was published [11], reeferred to as a Vision Transformer (ViT). The ViT decomposes an input image into patches, similar to words in a sentence, and processes them through multi-head attention. Additionally, a Multilayer Perceptron (MLP) is employed as the feedforward network. In later work Microsoft introduced a hierarchical division of image inputs and a shifted window approach in their Swin Transformer [16]. This was reported to outperform ViT by 2.4% in ImageNet-22K classification (21,841 different categories). Its version 2 [17] applied a cosine function in the attention module. enabling the scaling of capacity and resolution. More detail on transformer-based object detection is discussed in Section 3.4.2. To date, Swin Transformers have been widely adopted in a range of applications including image restoration [18].

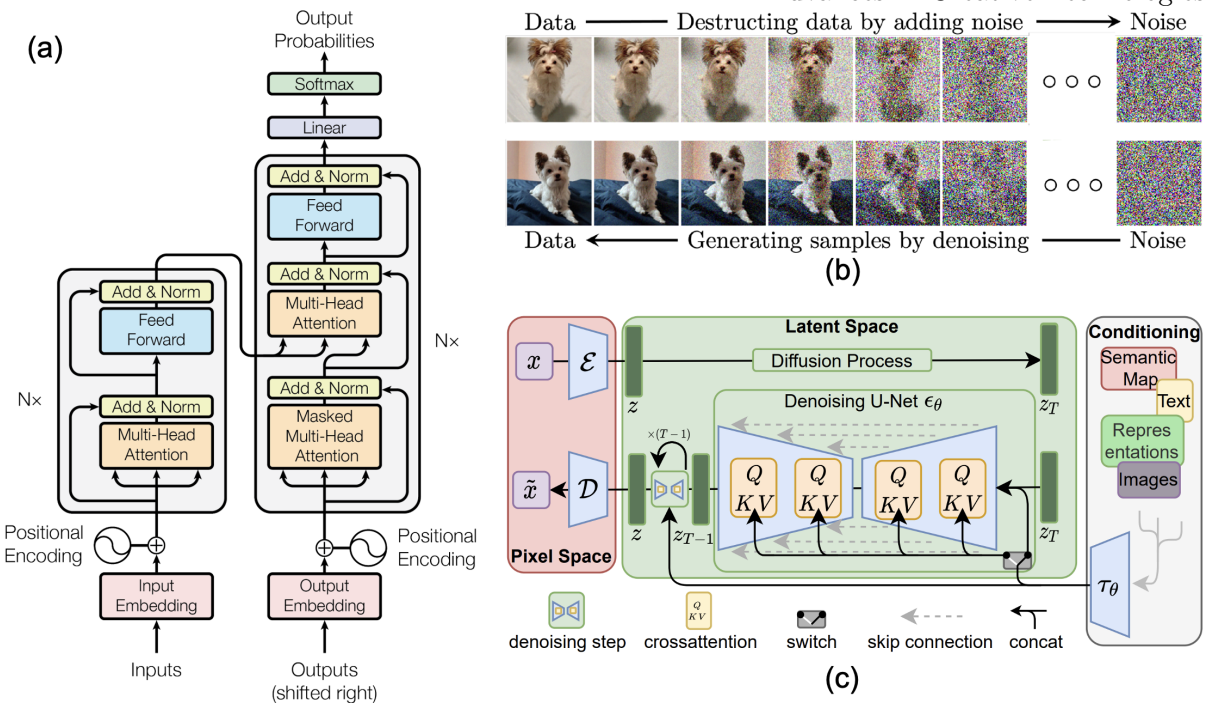

Figure 1: Generative AI. (a) Transformer architecture [10]. (b) The top row represents the diffusion process and the bottom row represents the generation process of the new image [23]. (c) Latent Diffusion Models (LDM) [24].

Comprehensive surveys on the use of transformers for image and video processing can be found in [19] and [20], respectively.

Transformers have been widely used and offer better performance across many tasks. One reason for this widespread adoption has been the availability of open-source Transformer libraries such as Hugging Face[8], a platform that assists developers to build applications for tasks including computer vision, NLP, audio, tabular data, multimodal tasks, and reinforcement learning. The platform also provides access to model zoo [9] pretrained networks and datasets.

In recent years state space models [21, 22], commonly known as 'Mamba' have emerged. These are a linear variant of Transformersdistinguished by their linear complexity in attention modeling. They are acknowledged to offer an equivalent or better performance than traditional Transformers, while demanding fewer computational resources and less memory.

2.2 Large language models

LLMs are based on transformer models using self-attention mechanisms as their core modules. Training LLMs comprises two steps: i) pre-training in an unsupervised learning manner, and ii) fine-tuning to a specific task or prompt-tuning for better user inputs. The models are first 'pre-trained' with a large amount of unlabelled text data to learn the meaning of words, and the relationships between those words, before using it to adapt to a downstream task. This is

[8] https://huggingface.co/
[9] Such as https://modelzoo.co/ and https://pytorch.org/serve/model_zoo.html

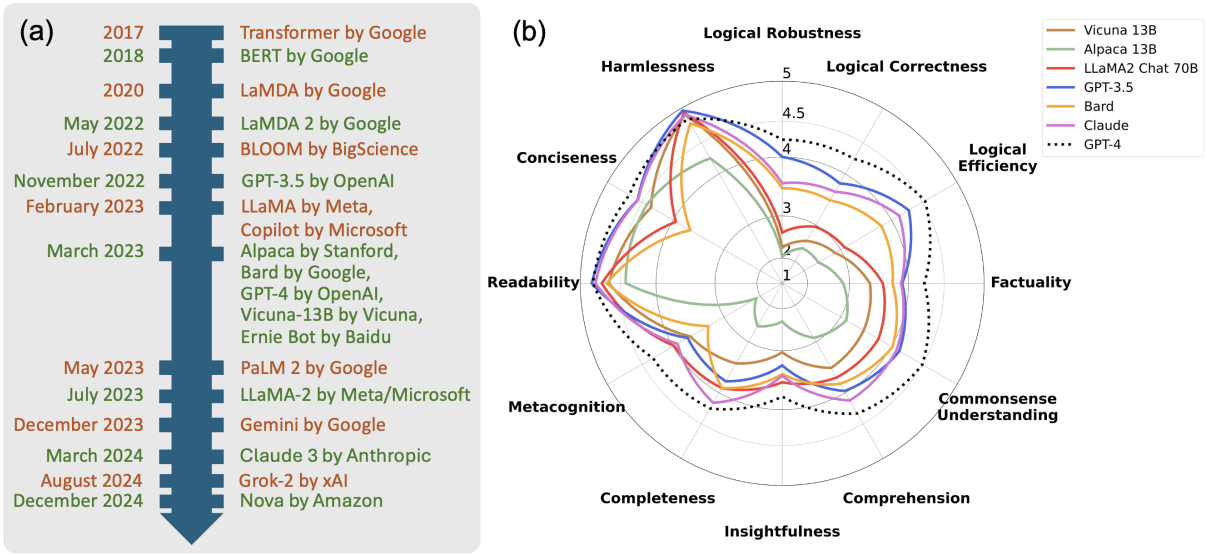

Figure 2: (a) Timeline of large language models. (b) Performance comparison evaluated by FLASK [30].

why OpenAI refers to their model as a Generative Pre-trained Transformer (GPT).

Fine-tuning involves training the model on new datasets. The drawback is however that these data need to be large enough to ensure generalization to new tasks. Prompt-tuning and prompt engineering are relatively new disciplines for developing and optimizing prompts to efficiently use language models. Prompts guide the way AI models interpret and respond to user queries. Prompt engineering is the process of structuring text or phrasing that guide the model towards generating the desired output. This relies heavily on trial and error, and an understanding of how the model responds. Prompt-tuning, on the other hand, involves training a small set of parameters before utilizing the LLM, thus requiring a relatively small amount of new data. This approach essentially converts text inputs into task-specific virtual inputs, referred to as tokens, while the pre-trained LLM remains unchanged [25]. The main drawback of prompt-tuning is lack of interpretability. This paradigm has however extended to other domains, such as visual prompt tuning [26]. For a comprehensive survey of LLMs, please refer to [27].

To date, there are many LLM platforms as shown in Fig. 2. Fig. 2(a) shows their timeline. Many surveys and evaluations of LLMs are also available [27–29]. These include FLASK (Fine-grained Language Model Evaluation based on Alignment SKill Sets) [30] which evaluates LLMs based on 12 fine-grained skills for comprehensive language model evaluation: logical correctness, logical robustness, logical efficiency, factuality, commonsense understanding, comprehension, insightfulness, completeness, metacognition, conciseness, readability, and harmlessness. Evaluation results from FLASK are shown in Fig. 2 (b).

2.3 Diffusion Models

A generative model, in the context of AI, exploits machine learning to learn a probability distribution of the training data to generate new data samples. The very first models were based

on Autoencoders (AEs) that learn to encode input data into a lower-dimensional representation (latent space) and then decode it back to its original form. A specific type of AE, a variational autoencoders (VAE) [31], learns the latent space as statistical parameters of probabilistic distributions, leading to significant improvement of the generated results. Concurrently, Goodfellow et al. [32] introduced an alternative architecture known as a Generative Adversarial Network (GAN). GANs comprise two competing AI modules: a generator, which creates a sample, and a discriminator, which determines whether the received sample is real or generated. When comparing VAEs to GANs, VAEs exhibit greater stability during training, whereas GANs excel at producing realistic images. More details about AEs and GANs for creative technologies can be found in our previous review [8].

An important factor in driving the rapid growth of generative AI has been the development of diffusion probabilistic models (referred to as diffusion models (DMs)). The first DM was introduced in 2015 by Sohl-Dickstein et al. [33], using Nonequilibrium Thermodynamics. However, it took a further 5 years for DMs to generate desirable results: the era of DMs began with Denoising Diffusion Probabilistic Models (DDPMs) proposed by Ho et al. [34] in 2020 and Score-based diffusion models proposed by Song et al. [35] in 2021. These involve a simplified process using a denoising autoencoder to approximate Bayesian inference. In brief, the models leverage a diffusion process to learn a probability distribution of the input data. As the name suggests, the data is diffused by gradually adding noise at each iteration step as shown in Fig. 1 (b). A deep neural network (DNN) is then trained to remove this noise, called the denoising process or reverse process. Consequently, the trained model uses random noise to generate data with characteristics similar to those of the training samples. Comparing to GANs, DMs provide higher diversity samples [36] and a training process that is much more stable and does not suffer from mode collapse. DMs are however computationally intensive and require longer training times compared to GANs. The complexity can significantly reduced by training the DMs in latent space. Latent Diffusion Models (LDM) [24] use pretrained networks to convert images to feature maps, and perform training on a low-dimensional space. The diagram of LDM is shown in Fig. 1 (c).

Generating a synthesized sample at random might not be particularly useful, especially for creative industry applications. Therefore, conditional diffusion models have been proposed, supporting a wide range of applications such as text-to-sound, text-to-images, and image-to-videos. For DMs, the conditional distributions are modelled using a conditional denoising autoencoder. Classifier guidance was introduced in [36] to improve the generation of images of a desired class. For example, when we provide the model with information, such as 'a flower', the DM will synthesize a variety of flower images, as the word 'flower' guides the model toward the latent distribution that is formed by various images of flowers. The work in [37] simply refines the latent space of well-trained unconditional DDPM so that the higher-level semantics of the synthetic samples are similar to the reference (conditioning). The LDM [24] offers more flexible conditional image generators by adding cross-attention layer (referred to Transformers in Section 2.1) to the denoising autoencoder. A survey on the methods and applications of DMs prior to 2024 can be found in [38].

2.4 Implicit Neural Representations

Implicit Neural Representations (INR), also called neural fields, neural implicits or coordinate-based neural networks, represent input content implicitly through learned functions F, as shown

in Eq. 3. They can be considered as fields x (represented by a scalar, vector, or a tensor with a value, such as magnetic field in physics) that are fully or partially parameterized by a neural network Φ, typically an MLP [39].

$$F(x, \Phi, \nabla_x \Phi, \nabla_x^2 \Phi, \dots) = 0, \quad \Phi : x \mapsto \Phi(x). \tag{3}$$

Although this concept appears complex, the process is actually very straightforward. For example, in the case of an image, the coordinates of each pixel (x, y) contain color information (r, g, b). The INR inputs (x, y) to the MLP and learns to provide the output (r, g, b). The weights and biases of the MLP now represent such an image. Usually, the number of parameters of the MLP is smaller than the total number of pixels multiplied by 3, accounting for the 3 color channels. Hence, one of its emerging applications is in data compression [40]. Moreover, the INR can handle complex and high-dimensional data efficiently, attracting attention for visual computing applications such as 3D scene reconstruction.

Traditional MLPs employ ReLU (rectified linear unit) for non-linear activation due to its simplicity. However, Sitzmann et al. [41] demonstrated that using periodic functions, such as sinusoids, are more suitable for representing complex natural signals, offering a better fit to the first- and second-order derivatives of the signals. However, this activation can cause ringing artifacts. Saragadam et al. instead proposed using complex Gabor wavelets [42], which learn to represent high frequencies better and simultaneously are robust to noise.

One of the fastest-growing areas that exploits INRs is **Neural Radiance Fields (NeRF)**, evidenced by 57 papers presented at CVPR, the largest annual conference in computer vision, in 2022 growing to 175 papers in 2023[10], before dropping to 71 in 2024, largely due to competition from 3D Gaussian Splatting[11]. First introduced in 2020 by Mildenhall et al. [43], NeRF is a form of neural rendering, a subset of generative AI, that generates novel views of a scene based on a partial set of 2D images. It achieves this by learning a mapping from 3D spatial coordinates and view directions (x, y, z, θ, ϕ) to colors and density (r, g, b, σ). This implicit representation allows NeRF to handle complex scenes with varying geometry and appearance, resulting in highly realistic renderings that include accurate lighting, shadows, and reflections. More detail can be found in Section 3.5.2.

3 Advanced AI for the creative industries

Similarly to our previous (2021) review of AI for the creative industries [8], Table 1 categorizes applications and corresponding AI-based solutions. These areas are explored in more detail below.

3.1 Content creation

Content creation is a fundamental activity of artists and designers and the term '*AI art*' refers to artforms created with the assistance of an AI algorithms or entirely by an AI system. This can refer to various digital forms including images, texts, audio, and videos. The roots of AI art can be traced back to the 20th century, exemplified by AARON, a computer program initiated

[10] https://markboss.me/post/nerf_at_cvpr23/
[11] https://github.com/Yubel426/NeRF-3DGS-at-CVPR-2024

Table 1: Creative applications and corresponding AI-based methods mentioned in this paper

Application		Technology		
		Trans./Attn.[1]	Diffusion model[2]	INR
Creation	Text	[4, 10, 44, 45]		
	audio/music	[46, 47]	[48–50]	
	Image	[46, 51]	[24, 51–58]	
	Animation/video	[59–69]	[63, 67, 69–75]	
	3D/AR/VR	[76]	[77–80]	[80–82]
Information Analysis	Text categorization	[83–86]		
	Film analysis	[87–89]		
	Content retrieval	[90–95]	[96]	
	Intelligent assistants	[97]		
Content Enhancement and Post Production	Enhancement	[98–103]	[104–107]	[108]
	Style transfer	[109–111]	[112, 113]	[110, 114]
	Super-resolution	[103, 115–122]	[122–126]	[123, 125, 127–129]
	Restoration	[5, 18, 103, 115, 119, 130–141, 141–143]	[106, 126, 128, 144–147]	[148]
	Inpainting	[149–153]	[124, 128]	
	Fusion	[154–157]	[158]	
	Editing/VFX	[159]	[159, 160]	
Information Extraction and Understanding	Segmentation	[5, 161–168]	[169–171]	[172, 173]
	Recognition	[11, 16, 17, 166, 174–180]	[181–184]	
	Tracking	[185–193]	[194–196]	[197]
	3D Reconstruction	[166, 198–203]	[204–207]	[43, 203, 204, 208–217], [218–225]†
Compression	Image*	[226–228]	[229–232]	[233–236]
	Video	[237, 238]	[239]	[40, 240–247]
	Audio*			
Quailty Assessment	Image*	[248–250]		
	Video*	[251–255]		

[1] Trans./Attn. include transformers, mamba and CNN-based architectures that use attention module.
[2] Some diffusion models employ the transformer in their denoising autoencoders.

† These methods are based on explicit neural representations.

* It is noted that for some compression and quality assessment tasks, there are other dominant network architectures in existing works. For example, LLMs have been used for image and audio compression, and visual quality assessment. Many neural audio codecs are also based on VQ-VAE models.

in 1972 to autonomously produce paintings and drawings [256]. The practicality of AI art has been enhanced with advancements in deep learning, particularly GANs from 2014 and, more recently, transformers, DMs and INRs.

3.1.1 Text generation, script and journalism

In the era of LLMs, AI writing tools have been widely used to assist various writing tasks, including generation written articles, blog posts, essays, and reports. These tools go beyond mere grammar and spelling checks; they boast advancements enabling them to analyze the style and tone of written material, adding images, videos and tables, offering suggestions to enhance clarity, coherence, and overall readability [257]. Moreover, AI tools extend their utility beyond content generation by automating tasks like keyword generation, meta tags, and descriptions, thereby increasing search rankings using search engine optimization (SEO). Additionally, they support the process of publishing across multiple online platforms. Transformers have been used to generate image captions by combining information from the images with a word prefix or questions [44].

AI script generators serve as beneficial aids for writers, filmmakers, and game developers, offering inspiration, idea generation, and assistance in crafting entire scripts [2, 258]. Human-AI brainstorming is helpful and saves time [259]. Presently, there are numerous software and websites providing both free and paid script generation services. However, many of these tools are still constrained when it comes to longform creative writing. Dramatron, developed by Google [260], introduces hierarchical language generation, enabling the creation of cohesive scripts and screenplays spanning long ranges. This includes elements such as titles, characters, story beats, location descriptions, and dialogue.

As discussed earlier, chatbots are now powered by LLMs, effectively simulating human conversation. These fundamental LLMs are specialized for specific tasks. For instance journalist AI and blog AI writers[12] generate content with layouts suitable for print or online publication. Additionally, AI tools exist that are designed to detect AI-generated content (e.g., for checking for copyright), AI-writing styles, content originality and to ensure the naturalness and flow of articles. Undoubtedly, generative AI is reshaping the way artists and journalists operate. For an in-depth exploration of the impact and implications of these technological advancements on news organizations, refer to the survey conducted by Beckett et al. [261].

Generating text and scripts automatically can also be done through image and video inputs without text prompts (e.g., image captioning [262]) and with text prompts. These approaches are referred to as Vision Language Models (VLMs): multimodal models that learn from images and text. The most common and prominent models often consist of an image encoder, an embedding projector to align image and text representation, often via a dense neural network, and a text decoder stacked in this order. The most well-known technique is Contrastive Language-Image Pre-training (CLIP) [263]. More recent work in [45] scales up the vision vocabulary by incorporating new image features into the existing CLIP model, resulting in improved content understanding. A comprehensive survey of VLMs for vision tasks can be found in [264].

[12]For example, see https://tryjournalist.com/

3.1.2 Audio and music generation

Similar to language models, AI-based music generation has rapidly advanced due to unsupervised learning on large datasets and the use of transformers (see Section 2.2). Examples of such systems include MuseNet[13], Magenta Studio[14], and Musicfy[15]. These tools assist in music composition by learning complex musical patterns, predicting the next word or music note in a sequence, and mixing specified instruments. Moreover, AI tools can convert one type of sound into another, such as from whistling to violin or from flute to saxophone[16]. This capability is invaluable for artists who may not be proficient in playing all the instruments they wish to incorporate, saving both time and costs. In 2024, Suno has released a model capable of producing radio-quality music that can be created in 2 minutes[17]. Later, Udio [18], was launched. This offers a prompt to create lyrics and music with a maximum duration of 90 seconds, and also appears to have, at least some, awareness of copyright.

AI voice software changes vocalizations from one person to another, for example enabling users to train the model to convert other people's voices into their own, e.g. lalals[19], Kits[20], Media.io[21], etc. Certain software, such as Voice.ai[22], even offers real-time voice changing capabilities. The technologies behind this uses a transformer to learn voice features and patterns in mel-spectrogram form. For example, the framework proposed in [49] uses a DM-based method with a transformer backbone to turn text input into a mel-spectrogram using the vector quantized variational autoencoder (VQ-VAE) [265]. Next, this mel-spectrogram is transformed into a sound wave. Unlike a regular spectrogram, the mel-spectrogram is based on the mel-frequency scale, which offers higher resolution for lower frequencies. Voice style transfer often uses zero-shot learning (a model is trained to recognize classes or categories that it has never encountered during training) [47] or few-shot learning (a model trained with only one or a few examples per class) [266]. Stable Audio Open [50] introduces a text-conditioned generative model for non-speech audio, trained on Creative Commons licensed data, capable of producing state-of-the-art 44.1kHz stereo audio.

Another emerging AI technology application is in the field of spatial audio. In 2022, Apple Music revealed that, in just over a year, more than 80% of its worldwide subscribers were enjoying the spatial audio experience, with monthly plays in spatial audio increasing by over 1,000%[23]. With head tracking, this technology significantly enhances the immersive experience. Masterchannel has launched SpatialAI[24], claiming it to be the world's first spatial mastering AI. This processes audio files and returns an optimized track for streaming platforms, along with an individually optimized stereo version for traditional distribution. All these advancements leverage transformer-based technologies.

[13] https://openai.com/research/musenet
[14] https://magenta.tensorflow.org/studio
[15] https://musicfy.lol/
[16] See an example by Ummet Ozcan at https://www.youtube.com/watch?v=lI1LCfTx2lI
[17] https://www.suno.ai/blog/v3
[18] https://www.udio.com/
[19] https://lalals.com/
[20] https://www.kits.ai/
[21] https://www.media.io/online-voice-changer.html
[22] https://voice.ai/
[23] https://www.apple.com/uk/newsroom/2023/01/apple-celebrates-a-groundbreaking-year-in-entertainment/
[24] https://platform.masterchannel.ai/spatial

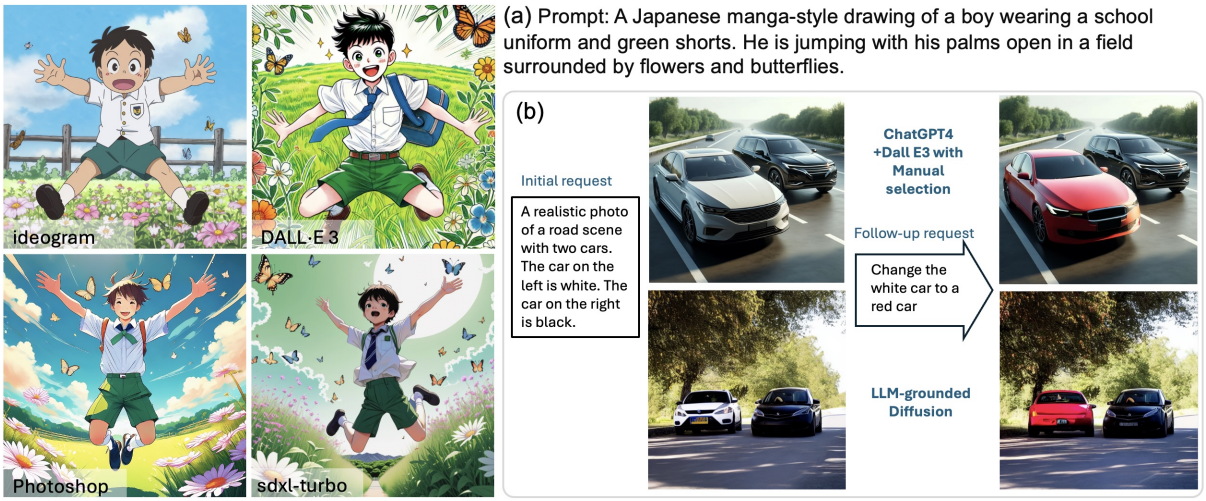

Figure 3: Text-to-image generation (generated on 27 November 2024). (a) text-to-image generation by Ideogram v1, DALL·E 3, Photoshop 2025, and sdxy-turbo by Nvidia. (b) The top-row images were generated by DALL·E in ChatGPT 4. The bottom-row images are generated by LLM-grounded Diffusion [55].

3.1.3 Image generation

As described in Section 2.3, recent advances in AI technologies for image generation are based on Diffusion Models (DMs). Well-known and highly competitive text-to-image models include Stable Diffusion[25], Midjourney[26], DALL·E[27], and Ideogram[28]. Released in March 2024, the latest version of Stable Diffusion (SD3), has been reported to outperform state-of-the-art text-to-image generation systems such as DALL·E 3 (released August 2023) [51], Midjourney v6 (released December 2023), and Ideogram v1 (released February 2024) in terms of typography and prompt adherence, based on human preference evaluations. These open-source tools are built on a Multimodal Diffusion Transformer (MM-DiT) architecture, which integrates attention from both text and images. LLM4GEN [58] fuses features from LLM and CLIP models to enhance the semantic understanding in text-to-image diffusion models, enabling them to better handle complex and dense prompts involving multiple objects. Examples of text-to-image generation are shown in Fig. 3 (a) comparing the performance of four models, i.e. Ideogram v1, DALL·E 3, Photoshop 2025, and sdxy-turbo by Nvidia. It is clear that hands are one of the most difficult features to generate, e.g., one hand has six fingers.

DALL·E 3, available on ChatGPT 4, also provides an inpainting tool, allowing the user to manually select the area to edit. However, as of April 2024, its performance is still limited. As illustrated in Fig. 3 (b), the selected area is the white car, and with the follow-up request to change the white car to the red car, DALL·E 3 generates correctly. However, if asked to replace it with a bicycle, it does not work. LLM-grounded Diffusion [55] was the first to introduce a framework that allows multiple rounds of user requests without the need for manual selection on the image. This is achieved by generating layout-grounded images, first using stable diffusion

[25]https://stability.ai/stable-image
[26]https://www.midjourney.com/home
[27]https://openai.com/dall-e-3
[28]https://ideogram.ai/

and then masking the latent variables as priors for the next round of generation[29]. Since then, text-driven image editing has seen significant improvements in quality, with most recent approaches adopting Diffusion Transformer architectures [57, 267].

Similar to DALL·E 3, Photoshop features a Generative Fill tool[30] designed to generate new images or assist with photo editing. It accepts a text prompt and provides several generation choices. After defining the editing area, users can remove and add new objects (more inpainting tasks are discussed in Section 3.3.5), transfer to new styles, and expand content within images. Recently, Brooks et al. introduced InstructPix2Pix [52], a conditional diffusion model that generates image editing examples without predefined editing areas. By combining GPT-3 and Stable Diffusion, the model effectively captures and matches the semantic meaning of the content in both text and image. Sometimes, style and context are not easy to describe in words. Textual Inversion [53] personalizes large pre-trained text-to-image diffusion models based on specific objects and styles, using 3-5 images of a user-provided concept. ByteDance announced Hyper-SD [56] which proposed trajectory segmented consistency distillation and provides real-time high-resolution image generation from drawing with a control text prompt.

3.1.4 Video generation and animation

Despite the success of text-to-image generation, text-to-video generation has not advanced at the same pace, only starting to grow more rapidly in 2024, largely due to its computational expense and content complexity. Several major companies and private platforms have however now released offerings, including Gemini 1.5 by Google, Make-A-Video by Meta, and Sora by OpenAI. Make-A-Video [70], through a spatiotemporally factorized diffusion model, leverages joint text-image priors and super-resolution in space and time. Some results however contain flickering artifacts[31]. Gen-2 by Runway[32] offers both text- and image-to-video and can generate a smooth 4-sec video. In April 2024, Adobe Premier Pro announced their integration of generative AI tools for video extension with third-party models by OpenAI, Runway and Pika Labs[33]. This new update also includes a contextual-selection tool, inpainting for object removal, and object addition to the defined areas in the videos with a text prompt.

Text-to-video technologies, combined with AI voice, have been tested not only by artists or producers but also by a wider audience. Results from these tests, such as automatically turning scripts into movie trailers and music videos, have been widely shared on public online platforms[34]. However, scene composition and transitions still require further editing to align with producers' needs[35]. In April 2024, Microsoft introduced VASA-1 [65], which turns a single static image and a speech audio clip into a video clip of realistic talking faces mimicking human facial expressions and head movements, as shown in Fig. 4 (right). The overall quality of the generated videos is better than VLOGGER by Google [66], which is based on similar technology – diffusion models. However, VLOGGER also offers movement of the upper body and hand gestures. Recently, ByteDance introduced an audio-driven interactive head

[29]Images in Fig. 3 (b) were generated using their demo: https://huggingface.co/spaces/longlian/llm-grounded-diffusion.
[30]https://www.adobe.com/th_en/products/photoshop/generative-fill.html
[31]https://makeavideo.studio/
[32]https://research.runwayml.com/gen2
[33]https://www.adobe.com/products/premiere/ai-video-editing.html
[34]https://twitter.com/minchoi/status/1775907105813217398
[35]See an example by Curious Refuge at https://www.youtube.com/watch?v=fJQbP34GoHQ

generation [69] that offers listening and speaking states during multi-turn conversations. This framework is based on a conditional diffusion transformer . The main technologies underpinning text-to-video and image-to-video tasks are based on DMs with a combination of 3D convolutions (or separately spatial and temporal convolutions), and spatial and temporal attention modules [72]. Tune-A-Video [73] modifies the style of an input video using a text prompt. The method leverages pretrained text-to-image models and introduces attention tuning to ensure temporal consistency. Early video generation methods often exhibit flickering, as observed in the CVPR2023 competition on text-guided video editing, where all results suffered from temporal inconsistency. Dreamix [71] videos do not have this issue, but they are very blurry. As an example of a transformer-based approach, CogVideo [59] employs VQ-VAE to convert input frames to tokens, which are then fused with text tokens to produce a new video. Phenaki [60] exploits transformers to generate variable length videos, but the quality is lower than those based on DMs. Evaluations of these methods can be found in [63]. More recent work has applied spatiotemporal layers to model temporal dynamics [67]. The transformer blocks have been redesigned for latent video diffusion modeling with window-restricted spatial and spatiotemporal attention. LaVie [74] demonstrates that simple temporal self-attention mechanisms, when combined with rotary positional encoding, are sufficient to capture the temporal correlations inherent in video data. The image-to-video generation process is analogous to text-to-video methods, but it conditions diffusion models on images rather than text. Some approaches further enhance generation by incorporating both textual descriptions (to guide motion) and images (to define objects and scenes) as inputs [75]. Many more free and commercial tools for video generation are now emerging. These include Veo 3 by Google DeepMind[36], Kling AI[37], Pika 2.2[38], Hailuo AI[39], etc. Though not perfect, the generated videos are close to reality (visit their websites for showcase examples).

Generating characters with human posture and motion from text prompts has also become popular. Make-An-Animation [61] trains on image-text datasets and fine-tunes on motion capture data, adding additional layers to model the temporal dimension. Animate Anyone by Alibaba Group [68] inputs a real photo or anime of a person with a sequence of guided poses. The results are significantly better than existing techniques, including Disco [64] and Bidirectionally Deformable Motion Modulation (BDMM) [62]. They also suggest using Animate Anyone with Outfit Anyone[40] to produce a character with a reference outfit.

Viggle[41] claims to be the first video-3D foundation model embodying an actual understanding of physics. It combines a character and a text prompt about motion to generate character animation. Available AI tools for 3D on the market include DeepMotion[42] that offers text-to-3D post animation and video-to-3D post animation, shown in Fig. 4 (left). The later function can track multiple people from real video and generates replicated characters with the same motions.

[36] https://deepmind.google/models/veo/
[37] https://www.klingai.com/
[38] https://pikartai.com/pika-2-2/
[39] https://hailuoai.video/
[40] https://humanaigc.github.io/outfit-anyone/
[41] https://viggle.ai/
[42] https://www.deepmotion.com/

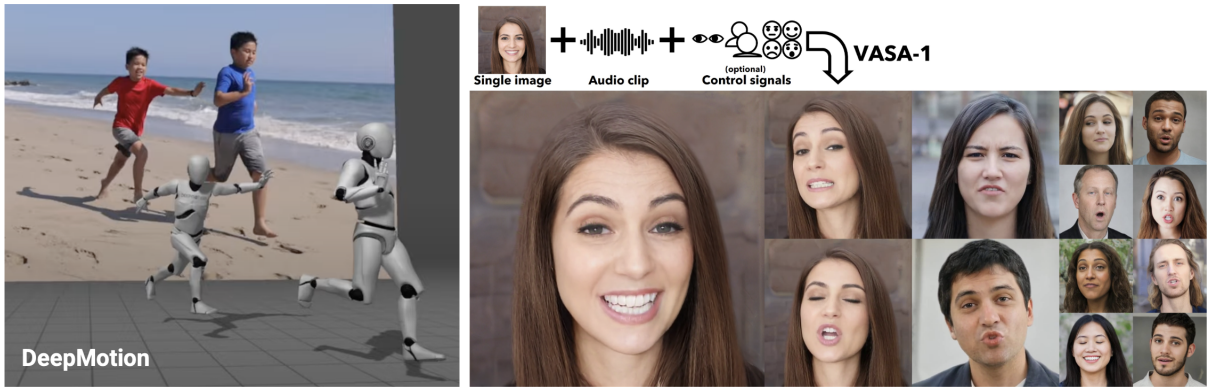

Figure 4: (Left) Video-to-3D post animation by DeepMotion. (Right) Image and audio to video by VASA-1 [65]

3.1.5 Augmented, virtual and mixed reality, and 3D content

While the benefits of LLMs in Augmented Reality (AR) directly target educational purposes, enhance cognitive support, and facilitate communication [268], mixed reality (MR) has once again become exciting since the release of the Apple Vision Pro in February 2024. This demonstrated the potential of MR experiences by merging real-world environments with computer-generated ones. Thanks to the rapid growth of AI-based 3D representation (see Section 3.5), the generation of AR/VR/MR content has advanced significantly. Real-time rendering with immersive interaction has improved, and real scenes can now be generated avoiding uncanny valley effects. There has also been an attempt to use autoregressive and generative models to estimate lighting, achieving a visually coherent environment between virtual and physical spaces in AR [82].

Similar to other content generation tools, LLMs have been influenced on immersive technologies, including text-to-3D and image-to-3D. Exciting examples include Holodeck [76], which automatically generates 3D embodied environments via text-prompt interactions with a large language model (GPT-4). 3D objects are gathered from Objaverse [269], a dataset with 800K+ annotated 3D objects. RealFusion [78], a single image to 3D object generator, merges 2D diffusion models with NeRF, improving Instant-NGP [209], which provides an API for VR controls. NeuralLift-360 [77] also uses diffusion models to generate priors for novel view synthesis. Magic123 [79] is the latest image-to-3D tool that uses 2D and 3D priors simultaneously to produce high-quality high-resolution 3D geometry and textures. DreamGaussian [80] offers text-to-3D and image-to-3D by adapting 3D Gaussian splatting (more in Section 3.5.3) into generative settings using a diffusion prior. This generates photo-realistic 3D assets with explicit mesh and texture maps within only 2 minutes. DreamGaussian4D [81] employs image-to-video diffusion and a 4D Gaussian Splatting representation to generate an image-to-4D model. The results are not very sharp, but they can be further edited with Blender.

In July 2024, Shutterstock launched its Generative 3D service in commercial beta, powered by NVIDIA Edify, a multimodal generative AI architecture. This service enables creators to rapidly prototype 3D assets and generate 360-degree HDRi backgrounds to light scenes using text or image prompts. In conjunction with OpenUSD, the created scenes can be rendered into 2D images and used as input for AI-powered image generators, allowing for the production of precise, brand-accurate visuals.

3.2 Information analysis

3.2.1 Text categorization

Applications of text categorization include detecting spam emails, automating customer support, monitoring social media for harmful content, etc. At its core, text categorization involves assigning predefined labels to text documents, which can be anything from a tweet to a lengthy article. LLMs are particularly well-suited for this task due to their ability to comprehend complex and nuanced language. One of the main advantages of using LLMs in text categorization is their transfer learning capability. Models can be pre-trained on a large amount of text and then fine-tuned on a smaller, task-specific dataset, with or without further post-processing technique. For example, CARP [83] applies kNN to integrate diagnostic reasoning process for final decision. ChatGraph, proposed by Shi et al. [84], utilizes ChatGPT to refine text documents. It uses a knowledge graph, extracted using another specific defined prompt, and finally, a linear model is trained on the text graph for classification. Multiple learners are also used to enhance the performances [85, 86].

3.2.2 Advertisements and film analysis

Not only does AI assist in generating ideas and content, but it can also aid creators in effectively matching content to their audiences, particularly on an individual level [270]. This effectively helps in advertising personalization—eMarketer[43] reported that nearly nine out of ten consumers are comfortable with their browsing history being utilized to create personalized ads. In contrast to outdated syntax-style searches, advanced LLM tools can comprehensively grasp user intent behind each search through conversation prompts, providing advertisers with a high level of granularity.

Current advances in generative AI would greatly benefit sentiment analysis, also known as opinion mining, where opinions are gather from social media, articles, customer feedback, and corporate communication and are analysed to understand emotion of the owners. This is a potential tool for filmmakers and studios, enabling the creation of effective and targeted marketing campaigns. By analyzing viewer emotions and opinions, AI can provide valuable insights into audience preferences, aiding in the optimization of film marketing strategies. Sentiment analysis with modern generative AI produce more accurate results. Technically, LLMs learn complex patterns and relationships in text data for sentiment classification [87, 88]. SiEBERT [89] provides pre-trained model with open-source scripts to be fine-tuned to further improve accuracy for novel applications. Cinema Multiverse Lounge [271], a multi-agent conversational system, allows users to interact with LLM-driven agents, each embodying a distinct film-related target users.

3.2.3 Content retrieval and recommendation services

Generative retrieval (GR) was pioneered by Metzler et al. [90]. Unlike traditional retrieval, which adheres to the "index-retrieve-then-rank" paradigm, the GR paradigm employs a single model to obtain results from query input. The model generally involve deep-learning based

[43] https://www.emarketer.com/content/spotlight-marketing-personalization

transformers, generating output token-by-token. More recent work in [95] introduces learning-to-rank training to enhance the performance system up to 30%. GR has several advantages including substituting the bulky external index with an internal index (i.e., model parameters), significantly reducing memory usage, and enabling optimization during end-to-end model training towards a universal objective for information retrieval tasks. Conversational question answering techniques have been integrated to enhance the document retrieval [94].

When retrieving visual content, recent work exploits generative models to enhance content-based model search [92]. These models decode the text, image, or video query into samples of possible outputs, which are then used to learn statistics for better matching between the query and output candidates. DMs are also employed for visual retrieval tasks, where they learn joint data distributions between text queries and video candidates [96]. A comprehensive survey on Generative Information Retrieval is available in [272].

While the retrieval task involves users directly defining a specific query input, recommendation services operate by retrieving content based on previous usage patterns. Essentially, a recommendation engine is a system that suggests products, services, or information to users through data analysis. Research in [273] has reported a positive association between buyers' attitudes toward AI and their behavioral intention to accept AI-based recommendations, with potential for further growth. Notable examples include the recommendation framework developed by Google [93], which utilizes GR. This framework assigns Semantic IDs to each item and trains a retrieval model to predict the Semantic ID of an item that a given user may engage with. A report by Aggarwal et al. [274] states that the recommendation accuracy of recommendation services has increased from 45.0% to 91.5% with the integration of generative AI.

3.2.4 Intelligent assistants

Intelligent assistants refer to software programs or applications that use AI and NLP to interact with users and provide helpful responses or perform tasks. These assistants can range from simple chatbots to sophisticated virtual agents capable of understanding and responding to complex queries. They're designed to assist users in various tasks, from answering questions and providing information to scheduling appointments and controlling smart home devices.

Current LLMs obviously enhance the performance of intelligent assistants, designed to understand complex inquiries and generate more natural conversational responses, such as Sasha [97]. Generative AI can also be used to enhance the performance of human customer support agents, aiding in search and summarization, as discussed in the previous section. Brynjolfsson et al. [275] examined the implementation of a generative AI tool designed to offer conversational guidance to customer support agents. Their research revealed that AI assistance significantly enhances problem resolution and customer satisfaction. Furthermore, they observed that AI recommendations prompt low-skill workers to adopt communication styles akin to those of high-skill workers. AI-based intelligent assistants may currently be more focused on educational purposes, but they can clearly help artists write more efficiently [276] or assist in customizing personal requirements [277]. The performance of personalized assistants can be enhanced with domain-specific knowledge to provide more in-depth responses to users [278].

3.3 Content enhancement and post production workflows

3.3.1 Enhancement

In our previous review paper [8], we discussed AI technologies for contrast enhancement and colorization as separate topics, as methods were developed specifically for each task. However, in recent years, there has been a shift towards addressing more complex issues, such as those encountered in low-light environments and underwater scenarios. These real-world situations often involve a combination of challenges, including low contrast, color imbalance, and noise.

In low-light conditions, scenes often exhibit low contrast, leading to focusing difficulties or the need for long exposures, which can result in blurred images and videos. To address this, LEDNet [279] has introduced a synthetic dataset for such scenarios and incorporated a learnable non-linear activation function within the network to enhance feature intensities. Meanwhile, SNR-Aware [98] estimates spatial-varying Signal-to-Noise Ratio (SNR) maps and proposes local and global learning branches using ResNet and transformer architectures, respectively. NeRCo [108] address low-light problem with INR, which unifies the diverse degradation factors of real-world scenes with a controllable fitting function. Diffusion models (DMs) have also become popular choices for low-light image enhancement [104–106]. Diff-Retinex [105] formulates the low-light image enhancement problem into Retinex decomposition, and employs multi-path generative diffusion networks to reconstruct the normal-light Retinex probability distribution. A recent state-of-the-art approach presented in [106] decomposes images into high and low frequencies using wavelet transform. High frequencies are enhanced using a transformer-based pipeline, while the low frequencies undergo a diffusion process. This method achieves nearly 2.8dB improvement over the state-of-the-art transformer-based approach, e.g. LLFormer [100], and significantly better than INR-based method, NeRCo [108], on a real low-light image benchmarking dataset. The technique has been extended for video enhancement in [107]. The output of the enhancement typically depends on user preferences. This has been viewed as a one-to-many inverse problem, with attempts to solve it using Bayesian approaches. For example, a Bayesian Enhancement Model (BEM) [280] incorporates Bayesian Neural Networks (BNNs) to capture data uncertainty and produce diverse outputs. The method can be used with Transformers or Mamba as the architecture backbone.

Regarding video enhancement, transformer and DMs are still in their early stages. STA-SUNet [101] has demonstrated that using transformers for low-light video enhancement outperforms CNN-based methods [281]. The recent Mamba-based network [282] also demonstrates promising results, outperforming STA-SUNet by more than 2 dB in PSNR. It is important to note that low-light enhancement is subjective. While most training datasets use normal lighting conditions as ground truth [283], the enhanced images and videos may alter the mood and tone of the content. Therefore, the tools for creative industries should be adjustable, not only for entire images and videos but also adaptive to specific areas and content. For instance, CLE Diffusion [129] enables user-friendly editing of lighting with fine-grained regional controllability.

Recent efforts have focused on enhancing User-Generated Content (UGC) videos—authentic recordings created by individuals rather than brands, often showcasing real experiences with products or services. The winning solution of the NTIRE 2025 Challenge on UGC Video Enhancement [284] implemented a pipeline of four sequential modules: color enhancement, denoising, BasicVSR++ restoration [285], and SwinIR [115]. This method achieved a 17% higher subjective score than the second-place entry, which used a two-stage framework, highlighting a

notable improvement in perceived visual quality.

3.3.2 Style transfer

Style transfer in AI art refers to a technique where the artistic style of one image (or video) is applied to another image (or video) while preserving the content of the latter. Style transfer has numerous applications in art, design, and image editing, allowing artists and designers to create unique and visually appealing compositions by blending different artistic styles with existing images (or videos). The applications also include image-to-image and sequence-to-sequence translations.

StyTr2 [109] is the first transformer-based method for style transfer, applying content as a query and style as a key of attention. InST [112] utilizes Stable Diffusion Models as the generative backbone and introduces an attention-based textual inversion module to learn the description of the content. StableVideo [113] uses a text prompt to describe the desired appearance of the output, transforming the input video to have a new look based on a diffusion model. For instance, a video of a white car driving in summer can be altered to show a red car driving in winter. A large pre-trained DM is employed in [111], where the style is injected to manipulate the self-attention of the decoder. To deal with the disharmonious color, they propose an adaptive instance normalization. A survey of style transfer using transformers and diffusion models can be found in [286]. Implicit Neural Representations (INRs) are less commonly used in style transfer tasks due to the difficulty of modeling the cross-representation between style and content. Moon et al. [110] combined INRs with vision transformers for generalizable style transfer; however, the results remain limited in quality. In contrast, the method proposed by Kim et al. [114] uses multilayer perceptrons (MLPs) to map image coordinates to the colors of the stylized output, guided by features extracted from both the content and style inputs to allow controllability.

3.3.3 Upscaling imagery: super-resolution (SR)

Impressive super-resolution (SR) results from transformer and diffusion models have been published extensively in the past few years. Originally, SR methods were developed using multiple low-resolution (LS) images, as different features in each image are combined to construct an enhanced one. However, these methods are not practical, as in most cases only one LS image is available. Hence, more methods have been developed for single image super-resolution (SISR).

The first use of a transformer, called ESRT, was for capturing long-term dependencies, such as repeating patterns in buildings. This was done in the feature domain extracted by a lightweight CNN module [116], outperforming those that use only CNNs. Since then, most SISR methods have been based on transformers. The Hybrid Attention Transformer (HAT) [118] was introduced, which improves the SR quality over ESRT by more than 2dB when upscaling 2×-4×. However, the NTIRE 2023 Real-Time Super-Resolution Challenge [287] showed that the winner, Bicubic++ [288], uses only convolutional layers and achieves the fastest speed at 1.17ms in upscaling 720p to 4K images. This method is significantly faster than any of the participants in the NTIRE 2025 Challenge [289], where Transformer-based architectures continue to dominate as the mainstream approach.

For DMs, SR3 by Google [123] has produced truly impressive results. It operates by learning

Figure 5: (Left) Examples of SR (×4) using generative model. (Right) Real-time portrait editing with FacePoke.

to transform a standard normal distribution into an empirical data distribution through a sequence of refinement steps, interpolating in a cascaded manner—upscaling 4× at a time. Later, IDM [125] combines INR with a U-Net denoising model in the reverse process of the DM. It is crucial to emphasize again that DMs are generative models. The SR results are generated based on the statistics we provide to the model during training (LR training samples). This is not for a restoration task, but rather for synthetic generation. A survey in SISR using DMs can be found in [290].

For video SR, numerous methods have emerged as part of a unified enhancement framework, as discussed in the previous section. One of the pioneering works to incorporate transformers specifically for video SR tasks is the Trajectory-aware Transformer for Video Super-Resolution (TTVSR) [117]. Although the results are slightly inferior to those of BasicVSR++ [285], which employs CNN and was introduced around the same time, both methods significantly enhance detail and sharpness compared to previous approaches, albeit not in real time. To address this limitation, the Deformable Attention Pyramid [291] has been introduced, offering slightly lower quality but a speed-up of over 3×. Recently, Adobe announced their VideoGigaGAN [121], which can perform 8× upsampling. This is achieved by adding flow estimation and temporal self-attention to the GigaGAN upsampler [120], which is primarily used for image SR, and text-to-image synthesis. Cao et al. [126] introduce a zero-shot video super-resolution framework that leverages a pre-trained image diffusion model, and replaces the spatial self-attention layer with a novel short-long-range (SLR) temporal attention layer. Recently, SeedVR integrated text information (captions) into a Diffusion Transformer (DiT) model, achieving state-of-the-art performance in video super-resolution.

Compared to traditional upscaling methods, generative AI can add details that did not exist in the original input image. These methods excel at generating high-quality natural images and structures, such as buildings, which are commonly included in training datasets. However, the process can be slow and may produce unpredictable results if the input image has very low resolution or contains content rarely seen in natural images. As shown in Fig. 5 (left), generative AI fails to upscale the knitting texture areas, instead generating lines more commonly found in typical images. While AI methods produce sharper edges, they perform less effectively on text.

3.3.4 Restoration

In our previous review paper [8], we categorized the work on restoration into several different types of distortions, including deblurring, denoising, dehazing, and mitigating atmospheric turbulence. Recent work however uses a unified network architecture to address these as inverse problems $y = hx + n$, where x and y are the ideal and observed data, respectively. h is a degradation function, such as blur, and n is additive noise. Often the super-resolution task is also considered as an inverse problem, meaning h includes downsampling process. Note that although designed as a single network, the model is trained with each distorted dataset separately.

The pioneering transformer-based method for image restoration, SwinIR [115], employs several concatenated Swin Transformer blocks [16]. SwinIR surpasses state-of-the-art CNN-based methods proposed up to the year 2021 in super-resolution and denoising tasks. The model is smaller and reconstructs fine details more effectively. Other two popular approaches that emerged in the same timeframe are Uformer [130] and Restormer [131]. Both incorporate Transformer blocks into hierarchical encoder-decoder networks, employing skip connections similar to those in U-Net. Their objective was to restore noisy images, sharpen blurry images, and remove rain. The networks focused on predicting the residual R and obtaining the restored image $\hat{x}$ through $\hat{x} = y + R$. While their performance is very similar, Restormer has half the parameters of Uformer. More recent, GRL by Li et al. [119] exploits a hierarchy of features in a global, regional, and local range using different ways to compute self-attentions as an image often show similarity within itself in different scales and areas. GRL outperforms SwinIR and Restormer. Additionally, Fei et al. introduced the Generative Diffusion Prior [128] for unsupervised learning, aiming to model posterior distributions for image restoration and enhancement. VmambaIR [143] incorporates Mamba blocks into the U-Net architecture, achieving superior performance compared to SwinIR and Restormer in both visual quality and model size.

For video restoration, the general framework comprises frame alignment, feature fusion and reconstruction. The process could be similar to image restoration but input multiple frames and run through the sequences in sliding window manner to exploit temporal information of a number of consecutive frames. Recently, Video Restoration Transformer (VRT) [103] and its improved version with recurrent process (RVRT) [99], have emerged as the state of the arts for video super-resolution, deblurring, denoising, and frame interpolation. This method introduces temporal reciprocal self-attention in the transformer architecture and parallel warping using MLP. These innovations enable parallel computation and outperform the previous state-of-the-art methods by up to 2.16dB on benchmark datasets. FMA-Net [102] proposed multi-attention for joint Video super-resolution and deblurring, achieving fast runtime with nearly 40% improvement over RVRT, and the restored quality was reported better by up to 3%.

For audio restoration, most software discussed in Section 3.1.2 offers tools for enhancing audio quality, such as eliminating background noise, echo, microphone rumble, and occasionally room reverberation, which have been well-established even before the advent of deep learning. There have been efforts to utilize AI for learning global contextual information to aid in the removal of unwanted sounds, leading to better final quality [134]. The latest advancements in this domain are primarily focused on addressing issues where significant portions of the audio data are missing. For instance, Moliner et al. [124] tackle problems such as audio bandwidth extension, inpainting, and declipping by treating them as inverse problems using a diffusion model. For a comprehensive survey on the use of diffusion models in restoration tasks, refer

to [292].

The following methods have been proposed for specific problems, but ideally, they should be adaptable for other tasks, even though they may not perform as well as they do for the original task.

i) **Deblurring**: A lightweight deep CNN model was recently proposed in [293], where a new discriminative temporal feature fusion has been introduced to select the most useful spatial and temporal features from adjacent frames. Feature propagation along the video is done in the wavelet domain. The deblurring performance is comparable to RVRT [99], but it is 5 times faster. DaBiT [133] mitigates focal blur content with depth information and applies SR for further enhancing fine details. Note that not only in software, but AI technologies have also been integrated into hardware. This includes autofocus, which is crucial for capturing sharp images of subjects, especially in dynamic environments where manual adjustments are impractical due to rapid movement. AI-driven autofocus methods have emerged, often tailored for specific camera hardware. For instance, Choi et al. proposed an autofocus model optimized for dual-pixel Canon cameras [294]. Additionally, Yang et al. investigated the correlation between language input and blur map estimation, utilizing semantic cues to enhance autofocus performance [132]. Remarkably, their model achieves comparable results to previous state-of-the-art methods while being more lightweight [295]. Autofocus could be used in conjunction with real-time object tracking (see Section 3.4.3) to produce desirable sharpness for moving objects in the video. Recently, Feng et al. [147] proposed a novel residual diffusion deblurring framework that integrates a conditional diffusion model guided by a defocus map and incorporates residual learning into the single-image defocus deblurring process.

ii) **Denoising**: SUNet [18] applies Swin transformer blocks combined in a UNet-like architecture. Denoising with diffusion models (DMs) [144] has been proposed by diffusing with estimated noise that is closer to real-world noise rather than Gaussian noise, achieving better performance than SwinIR [115] and Uformer [130]. INR with complex Gabor wavelets as activation functions show promising denoising results [42]. The NTIRE 2025 Image Denoising Challenge [296] revealed that the top-performing methods combined transformer-based and convolutional network architectures. Similarly, recent advances in video denoising also adopt a hybrid approach that integrates both architectures [141, 142].

iii) **Dehazing**: Vision transformers for single image dehazing was proposed in DehazeFormer [135]. Similar to SUNet, it is UNet-like architecture, but introduced Rescale Layer Normalization for better suit on improving contrast. The Fast Fourier Transform (FFT) has been employed in [140] due to the phase spectrum conveying more structural detail than the amplitude spectrum and demonstrating greater robustness to contrast distortion and noise. Then cross-attention between the RGB and YCbCr color spaces is applied. This approach achieves nearly 5 dB higher PSNR than DehazeFormer on a real-world smoke dataset. For video dehazing, Xu et al. [136] introduced a recurrent multi-range scene radiance recovery module with the space-time deformable attention. They also employs physics prior to inform haze attenuation. This method outperforms DehazeFormer by approximately 1dB.

iv) **Mitigating atmospheric turbulence**: Similar to dehazing, physics-inspired models have been widely developed to remove turbulence distortion [146, 148], while complex-valued CNNs have been proposed to exploit phase information [297]. There was also an attempt to use instance normalization (INR) to address this issue, providing solutions for tile and blur correction [148]. However, diffusion models have shown superior performance on single-image

restoration tasks [145], while transformer-based methods remained the state-of-the-art for video restoration [138, 139]. More recently, Mamba-based architectures have demonstrated their effectiveness in both visual quality and model efficiency [298]. A recent review can be found in [299].

3.3.5 Inpainting

Visual inpainting is the process of filling in lost or damaged parts of an image or video. CNNs and GANs have already achieved impressive results (see our previous review paper [8]). Recent work has focused more on editing rather than simply filling in the missing areas. This means users can now mask large areas of an image, and AI tools generate multiple results for users to choose from, a technique known as pluralistic inpainting [300]. Some notable methods include the following: Mask-Aware Transformer (MAT) [149] offers several outputs to fill a large missing area, consisting of a convolutional head, a transformer body, and a convolutional tail for reconstruction, along with a Conv-U-Net for refinement. PUT [150] proposes a patch-based vector VQ-VAE and unquantized Transformer to minimize information loss. Spa-former [153] employs a UNet-like architecture, where each level performs transformer with sparse self-attention to remove coefficients with low or no correlation, leading to memory reduction, while improving result quality by up to 5% compared to PUT.

Video inpainting presents greater complexity compared to image inpainting, despite the abundance of information available in an image sequence. The process typically involves tracking masks across frames, estimating optical flow, and ensuring temporal consistency. The current state-of-the-art methods include DLFormer [151] and ProPainter [152]. DLFormer conducts inpainting in latent space and utilizes discrete codes for video representation. On the other hand, ProPainter employs flow-based deformable alignment to enhance robustness to occlusion and inaccurate flow completion. The method excels in filling complete and rich textures, achieving a speed of 12 fps for full HD video. Video inpainting is also used for dubbing. DINet [301] replaces the mouth area to synchronize with a new language being spoken.

A comprehensive survey of learning-based image and video inpainting, covering approaches such as CNNs, VAEs, GANs, transformers, and diffusion models, can be found in [302]. Additionally, Elharrouss et al. [303] provide an in-depth review of the current challenges and future directions specific to transformer-based inpainting techniques.

3.3.6 Image Fusion

Image fusion is the process of merging multiple images from either the same source (such as varying focal points or exposures) or different modalities (e.g. visible and infrared cameras) into a single image. This process integrates complementary information from the various images to enhance overall quality, improve interpretation, and increase the usability of the final image.

Transformers and CNNs have been combined to extract global and local information, respectively. Most methods use CNNs for feature extraction, with transformers operating in the latent space [154,155]. Notable methods include SwinFusion [154], which utilizes a self-attention-based intra-domain fusion unit and a cross-attention-based inter-domain fusion unit to achieve multimodal and digital photography image fusion. Transformer-based image fusion has also been applied to downstream tasks like segmentation [156], achieving superior results by leveraging

the additional information. Self-attention blocks are employed to enhance intra-feature representations, while the cross-attention mechanism integrates inter-feature information to improve the quality of the fused output [157].

DDFM, the first diffusion model-based image fusion method, estimates noise in the reverse process by combining multiple inputs [158]. The expectation-maximization (EM) algorithm is integrated to estimate the noise distribution parameters, resulting in sharper images compared to traditional DDPM. For an in-depth review, the reader is referred to recent work in [304,305].

3.3.7 Editing and Visual Special Effects (VFX)

Editing or modifying specific areas of an image is much easier with DM technologies, particularly for headshot photos, such as targeting the eyes and mouth on the face [160]. This capability has been extended to video generation (see Section 3.1.4). Fig. 5 shows an example of the online tool, FacePoke[44], which allows users to move the head and modify the shapes of the eyes and mouth in real time. Motion-I2V [159] provides motion blur and motion drag tools to control specific areas of an image to add motion. The method is based on a diffusion-based motion field predictor and motion-augmented temporal attention.

VFX aims to create and/or manipulate imagery outside the context of a live-action shot in filmmaking and video production. When adding objects, scenes, and effects into traditional photographic videos, generative AI has obviously become an important tool, but some manual operations are still required. For example, in After Effects (EA)[45], the user selects the area where the object will be added and uses text prompts to describe such object. Subsequently, with the current EA version, the user will need to apply motion tracking so the generated object is moved accordingly.

AI technologies can upscale, enhance, and restore low-quality or old footage. For example, standard definition videos can be converted to high definition or even 4K quality without traditional manual remastering processes. This is particularly useful for remastering old movies or enhancing visual details in scenes. Generative AI has also simplified and accelerated automated processes, such as rotoscoping [306], an animation technique where animators trace over motion picture footage frame by frame to create realistic action. AI models can accurately detect and segment objects and characters in video frames, significantly speeding up the post-production process. Additionally, AI can assist the rapid creation of 3D models from 2D images generating realistic animations with minimal input data, facilitating complex human motions and synchronized facial expressions to voiceovers. One restriction is that current technologies still cannot yet generate full 4K accurate visual effects.

3.4 Information Extraction and Understanding

AI plays a crucial role in automating and optimizing the process of information extraction and understanding, enabling organizations to derive actionable insights from large and diverse data. Yan et al. [91] have categorized information extraction tasks based on the Format-Time-Reference space, as illustrated in Fig. 6 (a), where object detection and video object segmentation (VOS) are considered to be the simplest and the most complex tasks, respectively.

[44]https://huggingface.co/spaces/jbilcke-hf/FacePoke
[45]https://www.adobe.com/uk/products/aftereffects.html

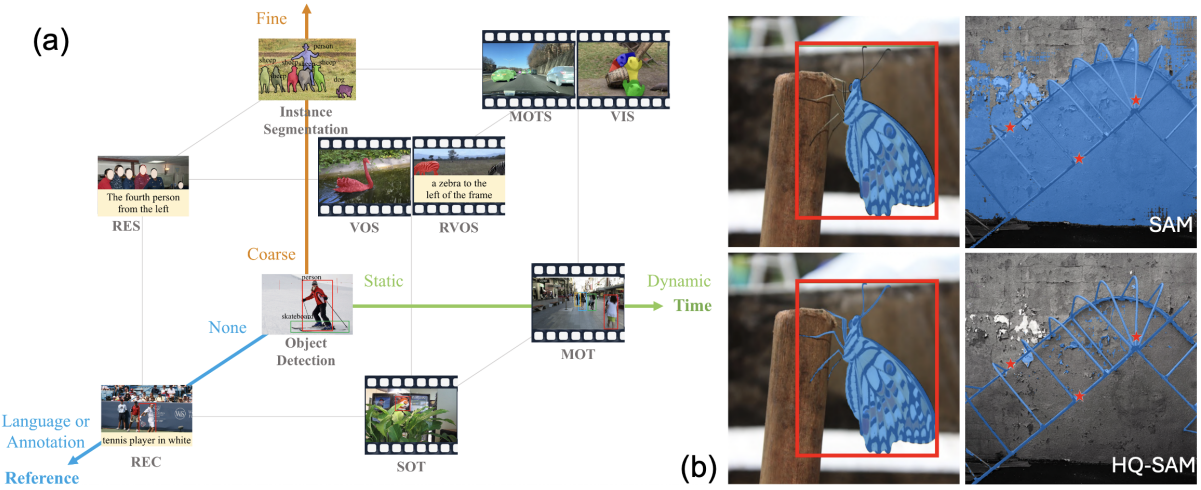

Figure 6: (a) Tasks in Object-centric understanding defined by Yan et al. [91] (REC:, Referring Expression Comprehension, RES: Referring Expression Segmentation, VOS: Video Object Segmentation, RVOS: Referring Video Object Segmentation, MOT: Multiple Object Tracking, MOTS: Multi-Object Tracking and Segmentation, VIS: Video Instance Segmentation, SOT: Single Object Tracking. (b) Current high-quality segmentation [163].

Recent advancements in this field draw significant inspiration from LLMs. These advancements include the utilization of prompts as conditional inputs for acquiring information. Moreover, following the pipeline approach used in LLMs, there is a growing trend towards leveraging very large datasets to pre-train models before fine-tuning them for specific downstream tasks. For instance, Meta AI [166] has introduced DINOv2, aimed at enriching information about visual content through self-supervised learning. This model was trained with 142 million carefully selected images, employing the ViT architecture. Google have introduced VideoPrism [178], a tool for scene understanding including classification, localization, retrieval, captioning, and question answering (QA). The model was trained on an extensive and diverse dataset consisting of 36 million high-quality video-text pairs and 582 million video clips accompanied by noisy or machine-generated parallel text.

3.4.1 Segmentation

The need for segmentation has grown dramatically in the past few years, given its central role in visual perception. Many segmentation methods now integrate an input prompt for users to define their preferred output appearances, such as pixel-wise segmentation, bounding boxes around objects, or segmented areas of interest. Most of these methods utilize transformer architectures [161]. Among them, Segment Anything (SAM) by Meta AI [162] stands out as a pioneer in promptable segmentation approaches. This method computes masks in real-time and has been trained with over 1 billion masks across 11 million images, facilitating transferability from zero-shot to new image distributions and tasks. HQ-SAM [163] enhances SAM by incorporating global-local feature fusion, leading to high-quality mask predictions. SegGPT [164] proposed context ensemble strategies and allows users to tune a prompt for a specific dataset, scene, or even a person, while SEEM [165] provides a completely promptable and interactive segmentation interface. More recently, SAM2 [167] introduced support for real-time video seg-

mentation. It is a unified model trained on a larger dataset than SAM. Interactive tools enable users to mark areas of interest and specify regions to exclude from the segmentation map. Zhou et al. propose an audio-visual segmentation (AVS) to generate pixel-level segmentation masks for sounding objects in audible videos. DVIS++ by [168] introduces a universal video segmentation framework capable of producing instance, semantic, and panoptic segmentation outputs. This transformer-based architecture comprises a segmentor, tracker, and refinement module, achieving state-of-the-art performance across several video segmentation benchmarks.

With DMs tehchnologies, Baranchuk et al. [307] have investigated semantic representation, and found DMs outperform other few-shot learning approaches. DiffuMask [169] automatically generate image and pixel-level semantic annotation using pre-trained Stable Diffusion with input as a text prompt. It has been proven that using these synthetic data improve segmentation accuracy. Currently, the state-of-the-art panoptic segmentation is the method developed by Nvdia, which is based on text-to-image DMs [170], outperforming the previous methods by up to 7.6%.

Applying INRs to segmentation is more popular in the medical domain, as the specific signals used, such as computed tomography (CT) and magnetic resonance imaging (MRI), can be formulated as continuous functions. In creative technologies, unsupervised domain adaptation (UDA) and INRs are used for continuous rectification function modeling in [172], achieving superior segmentation results in night vision. Recently, this work has been integrated with a non-local means block in [308] showning a significant improvement for instant segmentation in low-light scenes.

3D segmentation is also crucial for scene manipulation. In radiance fields, earlier segmentation methods required additional modules such as using k-means clustering to separate objects from the background [309]. However, the recent SA3D approach [173] segments 3D objects using NeRFs as the structural prior. SA3D operates by taking a trained NeRF and a set of prompts from a single view, then performing an iterative procedure. This involves rendering novel 2D views, self-prompting SAM for 2D segmentation, and projecting the segmentation back onto 3D mask grids. A comprehensive survey of 3D segmentation in computer vision can be found in [310].

3.4.2 Detection and recognition

Introduced in 2020, DETR by Facebook AI [174] was one of the first to adopt a transformer architecture for object detection. The approach achieves comparable results to an optimized Faster R-CNN [311], introduced in 2015. Deformable convolution has alson been used, (Deformable DETR [175]), resulting in training faster with approximately 5% accuracy improvement. A survey until 2022 [312] reported that Deformable DETR and Swin Transformers [16] outperform pure CNN-based YOLOv4 [313]. SwinV2 improves the first version by replacing original dot product attention with scaled cosine attention, improving accuracy by approximately 5%. Later, RT-DETR [314] improved inference speed by decoupling the intra-scale interaction and cross-scale fusion of features with different scales. RT-DETR is 25% faster than YOLOv8[46] with 6% improvement on MS COCO Object Detection dataset. Recently, YOLOv10 [315] has been released. YOLOv10 further improves the speed of detection approximately by 30% over RT-DETR with the same accuracy. A review of transformer-based methods for object detection

[46]https://github.com/ultralytics/ultralytics

can be found in [316,317]. Recently, YOLO12 [180] introduced an attention-centric architecture, achieving a 2.1% and 1.2% mAP improvement over YOLOv10-N and YOLOv11-N respectively, with only a slight decrease in speed.

To detect 3D objects, the transformer-based method MonoDTR [177] incorporates depth estimation from a single 2D image [201] to predict 3D bounding boxes. More 3D object detection methods have been developed for autonomous driving [318]; however, these approaches can also be adapted for AR and VR applications [179].

While DMs are primarily used to generate synthetic datasets [319, 320], they have also been demonstrated to function as zero-shot classifiers by Li et al. [181]. DMs are also of interest for detection tasks, Although feature extractors are still predominantly based on CNNs, such as ResNet, or Transformers (like Swin). DiffusionDet [182] formulates object detection as a denoising diffusion process from noisy boxes to object boxes, reporting performance that surpasses DETR. DMs have also been employed for anomaly detection [183, 184], functioning similarly to zero-shot classifiers.

3.4.3 Tracking

Object tracking stands out as one of the tasks that greatly benefits from transformers since *attention* is needed in both space and time. An experimental survey cited in [321] reveals that transformer-based methods consistently rank at the top of the leaderboard across various datasets. In the Visual Object Tracking (VOT) challenges of 2023[47], all of the top-10 employed transformer-based methodologies. The highest-performing approach achieved a 10% improvement in tracking quality compared to the winner in 2020. The current state-of-the-art for single-object tracking[48], however, is based on cross-attention and Mamba [193].

The first three tracking-by-attention approaches are TrackFormer [185], MixFormer [187], and ToMP [188]. TrackFormer extracts visual features using a CNN-based encoder, which are then tracked using a vanilla transformer [10] in a frame sequence, while MixFormer introduces cross-attention between the target and search regions. ToMP tracks the objects using prediction aspects. Many more methods have been proposed, including SeqTrack [190] and Track Anything Model (TAM) [189]. SeqTrack extracts visual features with a bidirectional transformer, while the decoder generates a sequence of bounding box values autoregressively with a causal transformer. TAM combines SAM [162] and XMem [322], offering tracking and segmentation performance on the human-selected target. However, the masked area is still not very sharp, and there is a subtle degree of temporal inconsistency. MOTRv2 [191] combines YOLOX [323] for object recognition and MOTR [186] for tracking, outperforming TrackFormer by 20%. Additionally, some methods have been specifically proposed for challenging environments, such as low light [192] and small objects, as seen in AnyFlow [197]. The latter exploits INR to upsample a continuous coordinate-based flow map, similar to SISR technique proposed in [127].

Similarly to detection tasks, DMs for tracking tasks are used as downstream processes by concatenating the diffusion head to the feature extraction backbone. However, a spatial-temporal fusion module has been added to the diffusion head to exploit temporal video features [194]. DiffusionTrack [195] localizes the target in a progressive diffusion manner, which is claimed to better handle challenging scenarios. The method in [196] exploits spatial-temporal weighting

[47] https://eu.aihub.ml/competitions/201#results
[48] https://paperswithcode.com/sota/visual-object-tracking-on-lasot

to suppress the probability of the tracker changing the target to the distractors. It, however, reports under-performance compared to MixFormer.

3.5 3D Reconstruction and Rendering

Bridging the gap between digital and physical realms, 3D reconstruction and rendering are integral to various creative technologies. In film and animation, they enable the creation of detailed digital models that blend seamlessly with live-action footage. Video games and digital twins leverage these technologies for dynamic environmental rendering. VR and AR use 3D reconstruction to create immersive and interactive experiences, with AR integrating digital content into real-world contexts. With recent AI technologies, 3D reconstruction and rendering have become faster and closer to reality. In particular, neural radiance fields and Gaussian Splatting enable artists and film producers to create shots that cannot be one in the real shooting environments.

3.5.1 Depth Estimation

Accurate depth information (alongside texture data) is typically required to construct 3D models. Depth sensors, such as lidar (Light Detection and Ranging) and structured-light 3D scanners, can be used for this purpose, but their applications are often limited by distance and cost. Consequently, vision-based sensors have become widely used. These sensors utilize two or more cameras to simulate human binocular vision or employ a single camera to capture images from different locations.

As deep learning can capture monocular cues such as object size, texture gradients, and perspective, depth estimation from a single image can produce accurate results. There have been attempts to use transformers, such as [199] and [200], and diffusion models, such as [205] and [207]. Amongst these, Depth Anything v2 [202] has become a state-of-the-art monocular depth estimation method. It is built on the previous version [201], jointly trained on large-scale labeled and unlabeled images and uses semantic priors from pretrained encoders. Depth Anything v2 significantly outperforms V1 in fine-grained details and robustness by using synthetic images and pseudo-labeled real images, as well as by extracting intermediate features from DINOv2 [166], which is trained with vision transformers. One of the notable capabilities of Depth Anything v2 is its ability to predict depth of transparent and reflective surfaces.

3.5.2 Neural Radiance Fields

Neural Radiance Fields (NeRFs), introduced in [43], have demonstrated the ability to learn a 3D scene from a smaller number of images captured from various viewpoints, as opposed to photogrammetry. They excel in neural rendering, particularly in view-dependent novel view synthesis, and have effectively tackled several challenges associated with automated 3D capture [39], such as accurately representing the reflectance properties of the scene. NeRFs offer high-resolution photo-realistic novel views and flexibility in postprocessing. They have hence gained significant attention in cinematography [324], as they offer reduced time and cost, particularly for outdoor shooting.

In the NeRF process (see Fig. 7 (a)), the camera positions and orientations are typically

estimated from a series of 2D images using techniques like feature-mapping and Structure-from-Motion (SfM), as demonstrated in [325]. Leveraging INR, each image (or camera pose) is mapped into camera rays that traverse the scene, generating 3D points with directional radiance (towards the camera). These points are then processed by an MLP to predict volume density and emitted radiance. Subsequently, volume rendering techniques are employed to generate an image, which is compared with the original via loss calculation. The MLP iteratively refines the model by minimizing this loss.

Since their introduction, there have been many variants of NeRFs aimed at improving their performance. Mip-NeRF360 [204] proposed unbounded anti-aliased technique achieving full 360 degree content. Google Research [210] trains NeRF from noisy RAW images captured in the dark scene, allowing changing viewpoint, focus, exposure, and tone mapping simultaneously. With segmentation techniques significantly advanced (see Section 3.4.1), there have been integrations utilizing semantic segmentation to enhance 3D representation [212]. DSEM-NeR [203] integrates the pretrained CLIP model to extract multimodal features—including color, depth, and semantics—from multi-view 2D images, thereby enhancing the reconstruction quality of complex scenes.

While the rendering quality of NeRF is very good, training and rendering times remain extremely high. The Instant-NGP tool developed by Nvidia [209] enables real-time training of NeRFs by bypassing sampling in empty spaces and dense areas, and by incorporating multi-resolution hash encoding techniques. These advancements substantially reduce the computational burden associated with representing high-resolution image features – training times have been reduced from hours to just a few seconds. Moreover, it offers VR controls for immersive 3D rendering experiences using OpenXR[49]. This allows users to navigate scenes, manipulate objects, and interact with the environment directly through VR headsets. Diffusion models are integrated to regularize NeRF reconstructions [206], resulting in smoother depth continuity and clearer edges where depth discontinuities occur.

The initial application of NeRFs to dynamic scenes was undertaken by Pumarola et al. [208], known as D-NeRF. However, the current leading method for generating high-quality novel views of real dynamic scenes is TiNeuVox [211]. It enhances temporal information by interpolating voxel features before feeding them into the radiance network to estimate density and color, similar to ordinary NeRF. DynVideo-E [214] adds an MLP to predict motion fields but focuses on human-centric content. PaReNeRF [217] addresses large-scale dynamic scenes using patch-based sampling. The main drawback of these methods is the large model size and/or long training time. Therefore, K-planes [218] propose a simple planar factorization for volumetric rendering, achieving low memory usage ($1000\times$ compression over a full 4D grid). Wavelet transform are employed in [215] to further reduce model size. KFD-NeRF [216] incorporates a Kalman filter-guided deformation field for more accurate motion estimation.

3.5.3 3D Gaussian Splatting

The main issue with NeRFs as a method to generate high-quality novel views is training time, which can exceed a day for high-resolution content on a single RTX 3090 GPU [213]. 3D Gaussian Splatting (3D-GS) [219] has been introduced to address this, using anisotropic 3D Gaussians to form a high-quality, unstructured representation of radiance fields. The process

[49] An open-source, royalty-free standard for access to virtual reality and augmented reality platforms and devices. https://www.khronos.org/openxr/

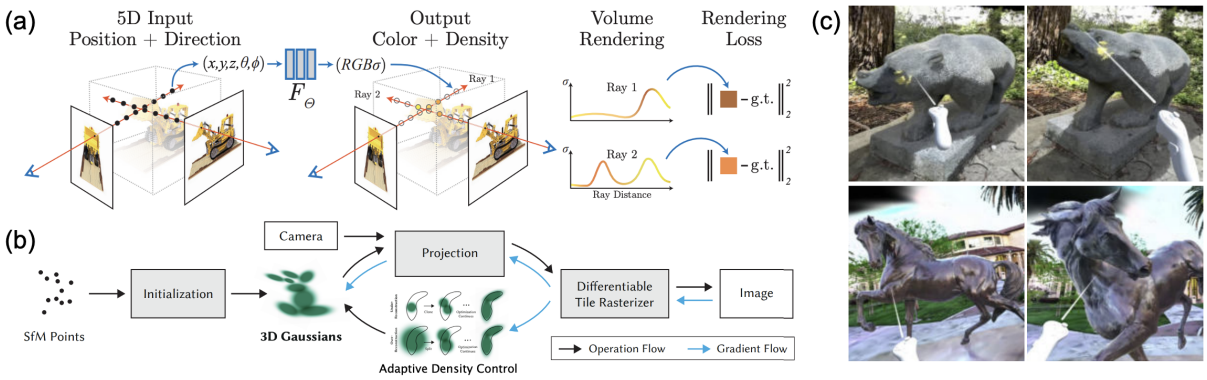

Figure 7: 3D representation. (a) Neural Radiance Fields (NeRFs) [43]. (b) Gaussian Splatting [219]. (c) Example scenes of VR-GS system for 3D content interaction in VR [326].

estimates a sparse point cloud through SfM. Each point possesses 3D Gaussian properties, such as position, covariance matrix, opacity, and spherical harmonics coefficients representing colors. The optimization of these parameters is interleaved with steps that control the density of the Gaussians to better represent the scene, as shown in Fig. 7 (b). A survey of 3D-GS can be found in [327].

In contrast to traditional NeRFs based on implicit scene representations, 3D-GS provides an explicit representation that can be seamlessly integrated with post-processing manipulations, such as animating and editing. VR-GS [326] offers intuitive and interactive physics-based gameplay with deformable virtual objects and realistic environments represented with 3D-GS. The example scenes are shown in Fig. 7 (c). Physics-inspired approaches are also integrated to improve 3D modeling in different media, such as 3D underwater scenes [223].

For dynamic scenes, 4D Gaussian Splatting (4D-GS) [220] introduces a Gaussian deformation field for motion and shape. It exploits a multi-resolution encoding method, achieving real-time rendering of up to 82 fps at a resolution of 800×800 pixels on an RTX 3090 GPU. Instead of developing in 4D, CoGS [221] exploits 3D-GS by integrating control mechanisms in separate regions to learn individual temporal dimensions. SC-GS [222] extracts sparse control points and uses an MLP to predict time-varying 6 DoF transformations. While the results show better visual quality than 4D-GS and CoGS, the performance heavily relies on camera pose estimation. Kong et al. [225] represent dynamic scenes using sparse, time-variant attribute modeling with a deformable MLP, while efficiently filtering out anchors corresponding to static regions. Their model achieves fast rendering speeds of over 110 FPS at a resolution of 960×540—nearly 10 times faster than SC-GS—and delivers a 1 dB improvement in PSNR.

LUMA AI[50] and Polycam[51] offer free tools for Gaussian splatting and photogrammetry creation for non-commercial use. The 3D objects created can be experienced with VR headsets for more immersive 3D and further used or developed in other applications. However, these tools have limitations in handling dynamic scenes due to occlusions, sparse observations per timestamp, and object reappearances over time. Rendering dynamic avatars can produce higher quality results by incorporating additional information. For example, EVA [224] disentangles the 3D Gaussian appearance into skeletal motion, facial expressions, body movements, and skin.

[50] https://lumalabs.ai/interactive-scenes
[51] https://poly.cam/captures

These components are then splatted to render the final photorealistic image.

3.5.4 Digital Twins

A digital twin is a virtual replica of a physical object, system, or process, continuously updated with real-time data for purposes such as simulation, testing, monitoring, and maintenance. This technology is increasingly adopted across various applications within the creative industries. For example, in product design and branding, it enables immediate observation of how a design performs in various contexts, facilitating the development of user-friendly products. Unilever reported that integrating digital product twins with 3D technologies, such as NVIDIA Omniverse[52], enabled the creation of product imagery twice as fast and 50% more cost-effective[53]. Digital twins also allow consumers to explore products or spaces virtually, simulating real-world interactions.

Accenture plc, a global professional services company, collaborated with Walt Disney Studios to develop digital twin technologies aimed at transforming the filmmaking process[54]. Their goal is to generate remotely accessible 3D models, enabling virtual exploration of potential shooting locations without requiring physical visits. The Virtual StudioLAB provides a digital replica created using 360-degree imagery and 3D modeling. These innovations have streamlined pre-production workflows for major productions from Marvel Studios and 20th Century Studios.

Digital representations such as avatars, proxies, and digital twins are increasingly being explored in artistic contexts, particularly in relation to identity, presence, and embodiment in virtual environments. The Tate Modern's film programme Avatars, Proxies and Digital Twins (Feb–May 2025) investigated these themes through curated audiovisual works, offering critical reflections on digital personhood. By engaging with diverse narrative forms, the programme highlighted the sociocultural implications of digital self-representation, prompting discourse on authenticity, agency, and the role of immersive media in shaping future human–machine interaction.

3.6 Data Compression

Data compression plays an important role in the delivery of creative content to audiences, effectively reducing memory and bandwidth requirements during signal storage and transmission [328]. Although coding methods based on conventional signal processing theories are still widely employed in most standards and application scenarios, learning-based solutions have emerged in research, showing great potential to achieve competitive performance in recent years. This subsection provides a brief overview of the recent advances in image, video, and audio compression, in particular focusing on the approaches proposed after 2021.

3.6.1 Image Compression

Since the first neural image codec [329] was proposed in 2016, numerous learning-based image compression methods have been developed, with significant performance improvements

[52] https://www.nvidia.com/en-gb/omniverse/
[53] https://www.unilever.com/news/press-and-media/press-releases/2025/unilever-reinvents-product-shoots-with-digital-twins-a
[54] https://www.accenture.com/mx-es/case-studies/communications-media/empowering-film-creatives-digital-twins

reported [330, 331]. Driven by the latest advances in neural network architectures, neural image codecs now outperform standard image codecs. Instead of using CNNs as the basic network structure, transformer-based architectures have become popular, offering the potential for better compression efficiency. Notable examples include SwinT-ChARM [226], STF [227] and LIC-TCM [228]. SwinT-ChARM [226] employs Swin transformers for non-linear transforms and outperforms the latest standard image codec, the Versatile Video Coding (VVC) Test Model (VTM, All Intra). STF [227] is based on a symmetrical transformer framework containing absolute transformer blocks in both the down-sampling encoder and the up-sampling decoder, which also shows improved rate-quality performance over VTM. LIC-TCM [228] exploits the local modeling ability of CNN and the non-local modeling performance of transformers, and proposes a parallel transformer-CNN mixture block. This new network structure, together with a channel-wise entropy model based on attention modules using Swin transformers, contributes to the superior performance of STF, with a more than 10% bitrate saving over VTM.

An alternative approach to learned image coding is based on advanced generative models. Early works [332,333] employed GANs to generate more photo realistic results with improved visual quality. Although these models fail to outperform conventional, CNN-based or transformer-based approaches, when distortion-based quality metrics, e.g., PSNR, are used for performance evaluation, they have been reported to perform well when perceptual quality models, such as MS-SSIM [334] and VMAF [335], or subjective tests are employed to measure perceived video quality. More recently, diffusion models have been applied in image compression to allow realistic reconstruction at ultra-low bitrates [229] achieving competitive performance compared to GAN-based models [230]. However, it should be noted that some of these generative models aim to generate (or synthesize) images with "perfect realism" rather than reconstruct results which are most similar to the original content. Notable work in this category includes image codecs using score-based generative models [231] and the diffusion-based residual augmentation codec (DIRAC) [232]. Moreover, another type of generative model based on INR has been employed for image compression; this learns a mapping between the spatial coordinates and the respective pixel values for the input image. The learned INR model is then compressed through parameter quantization and model compression to minimize the required bitrate. Notable INR-based image codecs include COIN/COIN++ [234, 235] and [236] that combine SIREN networks [233] with positional encoding.

In order to evaluate and compare neural image codecs under fair test conditions, public grand challenges have been increasingly run, typically associated with international conferences. One of the most well-known of these is the Challenge on Learned Image Compression (CLIC) [336]. In its latest competition, the best performing learned image codec [337], which is based on a GAN-enhanced Vector Quantized Variational AutoEncoder (VQ-VAE) framework, offered up to 0.6dB PSNR gain over VTM (version 22.2, All Intra) at similar bitrates; this codec is based on an autoencoder architecture with latent refinement and perceptual losses.

To support the deployment of neural image codecs, the International Organization for Standardization (ISO)/International Electrotechnical Commission(IEC) has developed a royalty-free learned image coding standard, denoted as JPEG AI [338], which aims to offer significant performance improvement over existing standards for both human and machine vision tasks. The Call for Proposals of JPEG AI was published in 2022, while the Working Draft and the Committee Draft outlining its core coding system were released in 2023 [339], with its first version published in October 2024 [339]. JPEG AI follows the same framework (the auto-encoder structure) as most existing neural image codecs, and its test model JPEG AI VM (version 4.3) has

been reported to achieve up to 28.5% coding gains over VVC VTM (All Intra mode) [340].

3.6.2 Video Compression

Compared to image coding, the compression of video content is a much more challenging task, particularly for immersive video formats and diverse content types. Although video coding standards including H.264/AVC (Advanced Video Coding), H.265/HEVC (High Efficiency Video Coding) and H.266/VVC (Versatile Video Coding) are still predominant in real-world applications, learning-based video coding has advanced dramatically in the past five years, with new deep learning enhanced conventional coding tools and end-to-end optimized neural video coding frameworks proposed.

i) **The enhancement of conventional coding tools** focuses on employing deep learning techniques to improve the performance of one (or multiple) coding modules in a standard-applicant codec. These modules include intra prediction [341], inter prediction [342], in-loop filtering [343], post filtering [344] and resolution re-sampling [345]. To facilitate efficient integration, the MPEG Joint Video Experts Team (JVET) built a test model in 2022 based on VTM 11, named Neural Network-based Video Coding (NNVC) [346], with its latest version NNVC-7.1 containing two major learning-based coding tools, neural-network based intra prediction and in-loop filtering, which has achieved an up to 13% coding gain over VTM 11 (Random Access mode) [347]. However, this learning-based codec requires much higher computational complexity (up to 477 kMACs/pixel) and high-spec GPU support compared to conventional codecs. Meanwhile, members of the Alliance of Open Media (AOM) have also developed multiple CNN-based coding tools for the next generation of video coding standard beyond AV1. The latest proposals focus on the trade-off between performance and complexity, with one of them based on inloop filtering and super-resolution, which achieves an average BD-rate saving of 3.9% (in PSNR) over AVM, the test model of AV2, but only requires a much lower computational complexity (below 1.5kMACs/pixel) [348]. More recently, research has been conducted to further improve the performance of these learning-based coding tools utilizing more advanced network architectures, including ViTs [349], and diffusion models [239]. There are also investigations on applying preprocessing before compression [350, 351], where the training of the deep preprocessors is based on proxy video codecs and/or rate-distortion loss functions to simulate the behavior of conventional video coding algorithms.

ii) **End-to-end optimized neural video codecs.** Alongside the enhancement of coding tools in conventional video codecs, more recent research activities have focused on using neural networks to implement the whole coding workflow, enabling data-driven end-to-end optimization. The performance of these neural video codecs has advanced significantly in the last five years, since the first attempt, DVC [352], was published. DVC matched the performance of a fast implementation of H.264 (x264). However currently, learned video coding algorithms (e.g., DCVC-FM [353] and DCVC-LCG [354]) are able to compete or even outperform the state-of-the-art standard codecs, such as VVC VTM under certain coding configurations. These learning-based methods often focus on enhancement from different perspectives, including feature space conditional coding (e.g., FVC [355] and DCVC [356]), instance adaptation [357,358], and motion estimation (e.g., DCVC-DC [359]). New architectures have also been proposed such as CANF-VC [360] based on a video generative model, MTMT [237] using a masked image modeling transformer-based entropy model and VCT [238] based on a video compression transformer. It is noted that although promising coding performance has been achieved in the

aforementioned works, these neural video codecs (in particular those based on autoencoder backbones) are typically associated with high computational complexity (especially in the decoder), which constrains their deployment for practical applications. To address this issue, researchers attempted to achieve complexity reduction while maintaining the coding performance through model pruning and knowledge distillation [361, 362].

It should be noted that the neural codecs mentioned above are typically trained offline with diverse video content [363], and deployed online for inference. In this case, model generalization becomes important, and this is why these codecs often have a large model capacity, resulting in large model sizes and slow inference runtime. Inspired by recent advances in implicit neural representations (INR), a new type of video codec has emerged that employs INR models to "represent" the video by learning a coordinate-based mapping and compressing the network parameters for transmission. This approach converts a video coding problem into a model compression task, which allows the use of a much smaller network to "overfit" the input video, with the real potential for fast decoding. Existing implicit neural video representation (NeRV) models can be classified into index-based and content-based methods. The former takes frame [240], patch [241] or disentangled spatial/grid coordinates [364] as model input, while content-based approaches [40, 242, 243] have content-specific embedding as inputs. Currently, one of the best INR-based video codecs [244] has already achieved a performance similar to that of VVC VTM (RA), but with a much lower decoding complexity compared to autoencoder-based neural codecs. Some of these models have also been applied to volumetric video content [246, 247], demonstrating their potential to compete with standard and other learning-based methods. However, it should be noted that the training of most NeRV models is based on an entire video sequence or even datasets; this results in a high system delay and does not meet the requirement of many low latency video streaming or real-time applications. To address this limitation, significant advances have been made [245] towards more practical INR-based video compression (such as the Low Delay and Random Access modes in VVC VTM [365]) by combining pre-training and online model overfitting.

Similarly to image compression, international grand challenges are used to compare neural video compression methods, with notable venues including the NN-based Video Coding Grand Challenge associated with The IEEE International Symposium on Circuits and Systems (ISCAS) and the Challenge on Learned Image Compression (CLIC, video coding track) with IEEE/CVF CVPR and Data Compression Conference (in 2024). The best performer in ISCAS 2024 NN-based Video Coding Grand Challenge offers an overall 55% BD-rate saving over HEVC Test Model HM [366], while the winner of the CLIC (video coding track) in 2024, a neural-network enhanced ECM codec [367] with a CNN-based in-loop filter, shows a more than 2dB (in PSNR) gain compared to VTM (RA) at the same bitrates.

3.6.3 Audio Compression

Similarly to images and videos, learning-based solutions have also been researched to compress audio signals, and most neural audio codecs are based on VQ-VAE [265]. SoundStream [368] is one of such models, which can encode audio content at various bitrates. It is based on a residual vector quantizer (RVQ) which trades off between rate, distortion, and complexity. This work has been further enhanced with a multi-scale spectrogram adversary and a loss balancer mechanism, resulting in improved rate-distortion performance. A more advanced universal model has been further developed [369] based on improved adversarial and reconstruction losses, which can

compress different types of audio. RVQ has also been extended from a single scale to multiple scales [370], which performs hierarchical quantization at variable frame rates.

More recently, researchers have started to exploit the use of LLMs for audio compression, leveraging the audio generation/synthesis abilities of generative models. UniAudio 1.5 [371] is one of such attempts, which converts an audio into the textural space, which can be represented by a pre-trained LLM that shares a similar backbone of UniAudio [372], a universal audio foundation model. LFSC is another neural audio codec based on LLMs, which achieved fast LLM training and inference through finite scalar quantization and adversarial training.

3.7 Visual Quality Assessment

Assessing the quality of visual signals remains an important and challenging task for many image and video processing applications. While subjective tests involving human participants remain the gold standard, objective quality models are frequently used because of their time and cost efficiency. These quality assessment methods are typically used to evaluate the performance of different visual processing approaches, and they can also be converted to loss functions, which are employed for optimizing learning-based processing models.

In recent years, quality assessment methods have been enhanced using deep learning techniques. The resulting learning-based quality models can quickly adapt to a specific type of content, leading to better performance compared to conventional, hand-crafted quality metrics. This section provides a brief summary of existing works in this research area, and highlights the main challenges which should be addressed in the near future. A more comprehensive overview of the image and video quality assessment literature can be found in [373–375].

3.7.1 Quality assessment models

Image and video quality assessment methods can be classified into two primary categories according to the availability of the corresponding reference content to the distorted test version: full-reference and no-reference models[55]. Prior to the AI era, conventional visual quality methods often exploit different characteristics of the human vision system and capture relevant information related to structural similarity (such as in SSIM and its variants [376–378]), distortion [379–381], and artifacts [382–384]. In many cases, the extracted features are further processed by models that simulate texture masking [385], contrast sensitivity [386], and saliency [387]. These hand-crafted quality models have also been combined with features within a regression-based framework in order to achieve more accurate prediction performance - VMAF is one such example [335]. When neural networks are involved for feature extraction, they are trained to capture information which can directly contribute to quality prediction through an end-to-end optimization strategy. Initially, convolutional neural networks were typically used, with notable examples such as DeepQA [388], LPIPS [389] and CONTRIQUE [390] for image quality assessment, and TLVQA [391], C3DVQA [392] and DeepVQA [393] for video quality assessment. Recent works have been reported to achieve better performance when Vision Transformers (ViTs) (or similar variants) are employed due to the effectiveness of their self-attention mechanism. Important works in this class include IQT [248], TRes [249], SaTQA [250],

[55]Reduced-reference quality metrics do existing in the literature, but the research in this field is less active in recent years.

FastVQA [251] and RankDVQA [252]. The former has been further extended as DOVER [253] and COVER [254] when aesthetic and/or semantic aspects in the content are taken into account.

More recently, inspired by the success of large language models (LLMs) [4, 394] in other machine learning tasks, these have been utilized in image and video quality assessment, demonstrating significant potential to achieve better model generalization. Q-Bench [395] is one of the first attempts that employs multimodal large language models to predict the perceptual quality of images based on prompt-driven evaluation. It queries the LLMs to provide information related to the final quality rating of the input image and the quality description. This has been further extended for video quality assessment tasks in Q-Align [396]. Other notable works include X-iqe [397] that performs the quality prompt in a multi-iteration manner focusing on both image fidelity and aesthetics. Prompt-based approaches have also been proposed for differentiating the quality difference between multiple images, such as 2AFC-LMMs [398] based on a two-alternative forced choice prompt and MAP (maximum a posteriori) estimation. Moreover, recent research works also focus on using pre-trained vision-language models, such as CLIP [263], which align better image and text modalities. Important examples in this class for image quality assessment include ZEN-IQA [399], QA-CLIP [400] and PromptIQA [401]. Similar works have also been proposed for video quality assessment, such as BVQI [402, 403] and COVER [254].

To support the training and validation of learning-based quality assessment models, image or video databases containing ground-truth subjective quality scores are typically employed. Commonly used image quality databases include LIVE [404], CSIQ [380], TID2013 [405], PieAPP and PIPAL, while video quality databases such as LIVE-VQA [406], KoNViD-1K [407], YouTube UGC [408] and LIVE-VQC [409] are typically employed for benchmarking in the literature. There are also databases developed that investigate the impact of specific video formats and/or artifacts, such as LIVE-YT-HFR [410] focusing on frame rates, VSR-QAD [411] on spatial resolution (or super-resolution artifacts), BAND-2k [412] on banding artifacts and Maxwell [403]/BVI-Artifact [413] containing multiple artifacts commonly produced in video streaming. Based on these databases, many learning-based quality assessment models are trained to minimize the difference (L1 or L2 norm) between predicted quality indices and subjective scores. However, due to the limited number of ground-truth quality labels associated with these databases and the resourcing-costing nature for collecting subjective data through human participants involved in psychophysical experiments, this type of training methodology cannot offer satisfactory performance, in particular when the model capacity is large. Moreover, since the experimental settings and conditions used for quality labeling are different in these databases, intra-database cross-validation is always required due to the limited model generalization and potential overfitting problems.

To address these issues, various proxy quality metrics have been used to label images and videos, which avoid expensive subjective tests and enable the generation of a large amount of training material with pseudo-ground-truth quality annotations. To further improve the reliability of quality labels, instead of learning the absolute values of the quality labels, ranking-inspired training strategies have been developed, which focus on improving the monotonicity characteristics of quality. Important examples based on these weakly supervised training methodologies include RankIQA [414] and UNIQUE [415] for the image quality assessment task, and VFIPS [416] and RankDVQA [252] for video quality assessment. Moreover, different self-supervised learning approaches have also been employed, which transform quality labeling to an auxiliary task. For example, CONTRIQUE [390] learns relevant features from an unannotated

image database based on the prediction of distortion types and degrees through contrastive learning. This method has been further applied to video quality assessment, resulting in a contrastive video quality estimator, CONVIQT [417]. More recently, quality-aware contrastive loss has been designed in [255, 418] to stabilize the learning process.

3.7.2 Performance and main challenges

Due to the lack of standard test conditions and limited model generalization within many existing image and video quality assessment models, deep compression methods are typically trained and benchmarked using different databases in conjunction with intra-database cross-validation. This can result in inconsistent evaluation results and conclusions. To enable a fair and meaningful comparison, various challenges and contests have been held for visual quality assessment. The Sixth Challenge on Learned Image Compression (CLIC) [336] associated with the Data Compression Conference 2024 is one of the latest examples which includes two quality assessment tracks for image and video compression. The best performer in the video quality assessment track achieves a Spearman Ranking Correlation Coefficient value of 0.825 [252], which is based on a ranking-inspired training methodology. Other notable challenges include the IEEE/CVF WACV 2023 HDR VQA Grand Challenge and the Video Super-Resolution Quality Assessment Challenge in ECCV 2024, which focus on high dynamic range and super-resolved content, respectively.

Although significant progress has been made in the past few years in visual quality assessment, including new models and training methodology, challenges remain, including limited model generalization and high computational complexity.

Another important use of quality metrics is as embedded loss functions for image and video processing optimization. This requires further capability and robustness, alongside complexity reduction, all topics to be addressed in future work.

4 Closing Thoughts and Future of AI in Creativity

This paper has presented a comprehensive review of current AI technologies and their applications that have emerged in recent years. Generative methods have driven a rapid growth in AI usage, particularly in the creative sector, significantly advancing the state of the art across various creative applications such as content creation, information extraction and analysis, content enhancement and data compression.

Through these applications, generative AI has not only broadened creative possibilities but has also reduced the manual effort and time traditionally associated with the production pipeline, allowing for greater creative experimentation and quicker production cycles. As this technology advances, it promises to unlock even more sophisticated capabilities in the creative industry. However, creative technologists, artists and other users must adapt, learn to use, and build these tools effectively and safely.

4.1 Challenges for AI in the Creative Sector

One of the primary challenges for artists engaging with modern generative AI and LLMs is the lack of consistent, controllable output. These models operate via stochastic sampling from high-dimensional latent spaces, meaning that identical prompts can yield different results across runs. This unpredictability makes it difficult for artists to achieve and iterate toward a precise creative vision. Although prompt engineering has emerged as a technique to guide model behavior, it requires technical knowledge and iterative refinement, which may not align with the intuitive or exploratory approaches common in artistic practice.

Moreover, there is a fundamental tension between the structured nature of current AI pipelines and the nonlinear, often improvisational workflows of creative disciplines. Many generative tools were originally designed for tasks like software development, content automation, or optimization [419], and are ill-suited for open-ended, exploratory creation. Artists typically work in cycles of ideation, experimentation, and revision—processes that demand fluid, real-time interaction and control, which existing AI systems struggle to support. These limitations point to a gap in current AI design: a need for systems that not only generate high-quality content but also adapt to the iterative, interpretive nature of artistic production. One possible approach addressed to these challenges is a shift toward top-down creative workflows, where artists define high-level concepts, themes, or goals via text prompts before refining specific outputs. This approach helps align AI-generated results with artistic intent, offering a degree of control over inherently stochastic systems.

Speaking at the World Government Summit in Dubai in 2024,[56] NVIDIA CEO Jensen Huang argued that with rapid advancements in AI, learning to code may become less essential for newcomers to the tech sector. He envisioned a future where traditional programming could be replaced by more intuitive AI-driven tools, thereby automating complex tasks and enhancing productivity—particularly for artists without coding expertise. While this perspective remains debated, it highlights the potential for AI to become more accessible within creative fields. However, AI-assisted coding tools are insufficient for creative practitioners, as artistic workflows tend to be unstructured and rely on domain-specific data. Artists must hence turn to techniques such as fine-tuning pre-trained models, few-shot learning, or domain adaptation—methods that are powerful yet typically inaccessible without machine learning expertise.

There is also broader concerns persist regarding the long-term impact of AI on the creative industries, particularly with the potential emergence of artificial general intelligence (AGI). Envisioned by organizations like OpenAI, DeepMind, and Anthropic, AGI could surpass human cognitive abilities, raising ethical and existential questions about the role of human agency in artistic expression.

4.2 Ethical Issues, Fakes and Bias

Advancements in generative AI, exemplified by models like Sora and Gemini 1.5 Pro, provoke ethical concerns and societal implications. These models, capable of generating highly realistic content, escalate the risk of misuse, through malicious deepfakes and misinformation. We are now in a situation where AI results transcend the uncanny valley, further complicating matters and challenging perceptions of authenticity. For example, the artist Miles Astray demonstrated

[56] https://blogs.nvidia.com/blog/world-governments-summit/

that even authentic photographs could be mistaken for AI-generated images. His real photograph 'F L A M I N G O N E' won both the jury's award and the people's choice award in the AI category of the 1839 Awards. His aim wasto highlight the ethical dilemmas inherent in AI, suggesting that the benefits of discussing AI's ethical implications could surpass the ethical concerns related to viewer deception[57].

While democratizing AI tools no-doubt presents opportunities to transform creative processes and workflows, it also necessitates robust regulatory frameworks to safeguard privacy and ownership. For example, deepfake technologies stimulate significant concerns about the spread of misinformation and other malicious uses. Efforts to detect and identify increasingly realistic deepfakes are thus as important as the generative methods used to produce them. These must however be accompanied by increased media literacy, and policies that address the ethical and legal implications.

Diversity and representation is a key issue when using AI tools. Unified Concept Editing [54] has been proposed as a basis for image generation in digital mediums. This aims to ensure the production of safe content with diverse representation, reducing gender and racial biases. Hallucination in generative AI (the production of outputs that are not faithful representations of reality but instead contain imagined or unrealistic elements) are a further cause of concern. These undermine trust in AI processes and can be due to limitations in the training data, biases in the model architecture or imperfections in the optimization process. Hallucinations associated with LLMs are one of the issues highlighted by the UK Government [1], alongside bias, regurgitation of private data, difficulties with multi-step tasks and challenges in interpreting black-box processes.

Governments across the world are increasingly expressing concerns about the challenges and uncertainties that generative AI technologies pose to rights holders and human creativity [2]. Generative AI presents substantial legal challenges, including the copyright status of AI-generated work and the intellectual property and copyright implications of the datasets used in training AI models. Viewpoints on this issue do however differ. For example, the track "Heart on My Sleeve," penned by an (as yet unidentified) human author, featured AI-generated vocals that replicated the voices of Drake and The Weeknd. Released independently on April 4, 2023, it was accessible via streaming platforms including Apple Music, Spotify, and YouTube. The song quickly became viral, accumulating over 20 million views across all platforms[58], prior to its removal by Universal Music Group, Drake's recording label. In contrast, Canadian artist Grimes has extended an invitation to musicians to emulate her voice via AI for the creation of new musical pieces, stipulating that the lyrics should not be harmful. She has advocated for the democratization of art and the abolition of copyright[59]. Additionally, Grimes has employed AI to design visual content for her LED backdrop at Coachella in 2024.

Finally, the rapid development of AI technologies has also raised concerns about job displacement and the balance between automation and human participation in creative processes. Ensuring that AI augments, rather than undermines, human effort poses a significant challenge for developers and policymakers.

[57] https://www.milesastray.com/news/newsflash-reclaiming-the-brain
[58] https://www.nbcnews.com/pop-culture/viral-ai-powered-drake-weeknd-song-removed-streaming-services-rcna80098
[59] https://www.bbc.com/news/entertainment-arts-65385382

4.3 The future of AI technologies

Several key technological issues remain which need to be addressed if AI is to deliver its full potential. These in particular relate to training data, computational complexity and their depth of reasoning or planning, and are discussed below.

A substantial amount of data is essential for training AI models in order to achieve high performance and good generalisation. Major companies such as Google, Meta, and NVIDIA, with their respective models: BERT, Segment Anything, and Canvas, dominate this space, benefiting from leveraged resources to gather data and process it to train sophisticated models. However, in November 2024, Bloomberg reported that OpenAI, Anthropic, and Google are all experiencing relatively slow growth in the performance of their AI models, with one of the key challenges being training data[60].

LLMs excel in applications involving complex tasks, advanced reasoning, data analysis, and understanding context. However, these models typically require high computational resources or cloud computing for development, operation and fine-tuning. A new trend emerging in alongside LLMs is the development of Small language models (SLMs), such as Phi-3 by Microsoft[61]. SLMs offer promising solutions for regulated industries and sectors encountering scenarios where high-quality results are essential while keeping data 'on-site'. Their potential is particularly relevant when deploying more capable SLMs on smartphones and other mobile devices, allowing them to operate 'at the edge' without relying on cloud connectivity. Recent highly successful platforms, such as DeepSeek-V3 [420] and Qwen2.5-Max [421], are based on Mixture-of-Experts (MoE) models, which tackle complex problems by dividing them into simpler sub-tasks, each handled by a specialized "expert."

Despite the evident advancements in AI, current models still struggle with tasks requiring planning or deep reasoning and are prone to errors when encountering unexpected data. This, in turn, reduces the confidence of users and trust in the results. AI algorithms can learn through reinforcement learning, but this process often identifies the best outcome as an anomaly rather than the norm. Yann LeCun, Professor at NYU and Chief AI Scientist at Meta, noted that while LLMs show a degree of comprehension in processing and generating text, their understanding lacks depth, often leading to results that defy common sense[62]. He advocates for self-supervised learning as a pivotal future direction for AI, emphasizing its potential to derive insights from unlabeled data. Concurrently, Andrew Ng, Adjunct Professor at Stanford University and Founder of DeepLearning.AI, sees iterative AI agentic workflows[63] as a key advancement for enhancing AI tool capabilities through an interactive approach by AI agents. These workflows involve autonomous agents that interactively learn from experience, understand natural language, and execute tasks on behalf of users.

The increasing openness of code and datasets is seen by many as a catalyst for accelerating AI advancements, with major firms like Microsoft, Google, and Meta supporting open access technologies. However, this openness also introduces security risks, necessitating new regulatory measures to monitor models post-release, to standardize documentation, and to assess the safety of software code and training data disclosure.

Finally, as stated in [2], the rapid advancement of AI technologies has revolutionized cultural

[60] https://www.bloomberg.com/news/articles/2024-11-13/openai-google-and-anthropic-are-struggling-to-build-more-advanced-a
[61] https://azure.microsoft.com/en-us/blog/introducing-phi-3-redefining-whats-possible-with-slms/
[62] https://twitter.com/ylecun/status/1728496457601183865
[63] https://www.youtube.com/watch?v=sal78ACtGTc

experiences, often referred to as 'CreaTech'—the convergence of the creative and digital sectors [422]. Such innovations not only reshape how people engage with art and creative work (e.g., through AR/VR/MR) but also drive the evolution of the technologies themselves.

References

[1] Communications and Digital Committee, "Large language models and generative AI," *1st Report of Session 2023–24*, 2024.

[2] L. Jeary and D. Gajjar, "Artificial intelligence and new technology in creative industries," October 2024, Accessed: 2025-01-10.

[3] Y. Bai, A. Jones, K. Ndousse, et al., "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv:2204.05862*, 2021.

[4] OpenAI, J. Achiam, S. Adler, S. Agarwal, et al., "GPT-4 technical report," *arXiv: 2303.08774*, 2023.

[5] X. Wang, W. Wang, Y. Cao, C. Shen, and T. Huang, "Images speak in images: A generalist painter for in-context visual learning," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6830–6839.

[6] N. C. Chung, "Human in the loop for machine creativity," in *AAAI Conference on Human Computation and Crowdsourcing*, 2021.

[7] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, and Y. Tang, "A brief overview of ChatGPT: The history, status quo and potential future development," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, 2023.

[8] N. Anantrasirichai and D. Bull, "Artificial intelligence in the creative industries: a review," *Artificial Intelligence Review*, vol. 55, pp. 589–656, 2022.

[9] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, et al., "On the opportunities and risks of foundation models," *arXiv:2108.07258*, 2021.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[12] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-maron, M. Giménez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, and N. de Freitas, "A generalist agent," *Transactions on Machine Learning Research*, 2022.

[13] G. Li, Q. Fang, L. Zha, X. Gao, and N. Zheng, "HAM: Hybrid attention module in deep convolutional neural networks for image classification," *Pattern Recognition*, vol. 129, pp. 108785, 2022.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[15] M.-H. Guo, T.-X. Xu, J.-J. Liu, and et al., "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, pp. 331–368, 2022.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[17] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11999–12009.

[18] C.-M. Fan, T.-J. Liu, and K.-H. Liu, "SUNet: swin transformer unet for image denoising," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2022, pp. 2333–2337.

[19] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, sep 2022.

[20] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapes, "Video transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12922–12943, nov 2023.

[21] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Conference on Language Modeling*, 2024.

[22] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," 2024.

[23] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Comput. Surv.*, vol. 56, no. 4, nov 2023.

[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10674–10685.

[25] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Nov. 2021, pp. 3045–3059.

[26] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision (ECCV)*, 2022.

[27] W. X. Zhao, K. Zhou, J. Li, T. Tang, et al., "A survey of large language models," *arXiv:2303.18223*, 2023.

[28] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, et al., "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, mar 2024.

[29] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, vol. 4, no. 2, pp. 100211, 2024.

[30] S. Ye, D. Kim, S. Kim, H. Hwang, S. Kim, Y. Jo, J. Thorne, J. Kim, and M. Seo, "FLASK: Fine-grained language model evaluation based on alignment skill sets," in *The Twelfth International Conference on Learning Representations*, 2024.

[31] D. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 2672–2680. Curran Associates, Inc., 2014.

[33] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning*, 07–09 Jul 2015, vol. 37, pp. 2256–2265.

[34] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, 2020, p. 6840–6851.

[35] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021.

[36] P. Dhariwal and A. Q. Nichol, "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[37] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "ILVR: Conditioning method for denoising diffusion probabilistic models," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14347–14356.

[38] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion models," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–20, 2024.

[39] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, "Neural fields in visual computing and beyond," in *Computer Graphics Forum*. Wiley Online Library, 2022, vol. 41(2), pp. 641–676.

[40] H. M. Kwan, G. Gao, F. Zhang, A. Gower, and D. Bull, "HiNeRV: Video compression with hierarchical encoding-based neural representation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 72692–72704, 2024.

[41] V. Sitzmann, J. N. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Proc. NeurIPS*, 2020.

[42] V. Saragadam, D. LeJeune, J. Tan, G. Balakrishnan, A. Veeraraghavan, and R. G. Baraniuk, "Wire: Wavelet implicit neural representations," in *Conf. Computer Vision and Pattern Recognition*, 2023.

[43] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., 2020, pp. 405–421.

[44] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "SIMVLM: Simple visual language model pretraining with weak supervision," in *International Conference on Learning Representations (ICLR)*, 2022.

[45] H. Wei, L. Kong, J. Chen, L. Zhao, Z. Ge, J. Yang, J. Sun, C. Han, and X. Zhang, "Vary: Scaling up the vision vocabulary for large vision-language models," in *European Conference on Computer Vision*, 2024.

[46] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. a. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, 2022, vol. 35, pp. 23716–23736.

[47] R. Huang, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "GenerSpeech: Towards style transfer for generalizable out-of-domain text-to-speech," in *Advances in Neural Information Processing Systems*, 2022.

[48] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," in *Advances in Neural Information Processing Systems*, 2022, vol. 35, pp. 4328–4343.

[49] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1720–1733, 2023.

[50] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Stable audio open," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[51] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, et al., "Scaling rectified flow transformers for high-resolution image synthesis," in *Proceedings of the 41 st International Conference on Machine Learning*, 2024.

[52] T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 18392–18402.

[53] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *The Eleventh International Conference on Learning Representations*, 2023.

[54] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzyńska, and D. Bau, "Unified concept editing in diffusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 5111–5120.

[55] L. Lian, B. Li, A. Yala, and T. Darrell, "LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models," *Transactions on Machine Learning Research*, 2024, Featured Certification.

[56] Y. Ren, X. Xia, Y. Lu, J. Zhang, J. Wu, P. Xie, X. Wang, and X. Xiao, "Hyper-SD: Trajectory segmented consistency model for efficient image synthesis," in *Advances in Neural Information Processing Systems*, 2024.

[57] K. Feng, Y. Ma, B. Wang, C. Qi, H. Chen, Q. Chen, and Z. Wang, "DiT4Edit: Diffusion transformer for image editing," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 3, pp. 2969–2977, Apr. 2025.

[58] M. Liu, Y. Ma, Z. Yang, J. Dan, Y. Yu, Z. Zhao, Z. Hu, B. Liu, and C. Fan, "LLM4GEN: Leveraging semantic representation of llms for text-to-image generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 5, pp. 5523–5531, Apr. 2025.

[59] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," in *The Eleventh International Conference on Learning Representations*, 2023.

[60] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, and D. Erhan, "Phenaki: Variable length video generation from open domain textual descriptions," in *International Conference on Learning Representations*, 2023.

[61] S. Azadi, A. Shah, T. Hayes, D. Parikh, and S. Gupta, "Make-An-Animation: Large-scale text-conditional 3D human motion generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 15039–15048.

[62] W.-Y. Yu, L.-M. Po, R. C. Cheung, Y. Zhao, Y. Xue, and K. Li, "Bidirectionally deformable motion modulation for video-based human pose transfer," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 7468–7478.

[63] Y. Liu, L. Li, S. Ren, R. Gao, S. Li, S. Chen, X. Sun, and L. Hou, "FETV: A benchmark for fine-grained evaluation of open-domain text-to-video generation," in *Advances in Neural Information Processing Systems*, 2023, vol. 36, pp. 62352–62387.

[64] T. Wang, L. Li, K. Lin, Y. Zhai, C.-C. Lin, Z. Yang, H. Zhang, Z. Liu, and L. Wang, "Disco: Disentangled control for realistic human dance generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 9326–9336.

[65] S. Xu, G. Chen, Y.-X. Guo, J. Yang, C. Li, Z. Zang, Y. Zhang, X. Tong, and B. Guo, "VASA-1: Lifelike audio-driven talking faces generated in real time," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[66] E. Corona, A. Zanfir, E. G. Bazavan, N. Kolotouros, T. Alldieck, and C. Sminchisescu, "Vlogger: Multimodal diffusion for embodied avatar synthesis," *arXiv:2403.08764*, 2024.

[67] A. Gupta, L. Yu, K. Sohn, X. Gu, M. Hahn, F.-F. Li, I. Essa, L. Jiang, and J. Lezama, "Photorealistic video generation with diffusion models," in *European Conference Computer Vision*, 2024, p. 393–411.

[68] L. Hu, "Animate Anyone: Consistent and controllable image-to-video synthesis for character animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 8153–8163.

[69] Y. Zhu, L. Zhang, Z. Rong, T. Hu, S. Liang, and Z. Ge, "INFP: Audio-driven interactive head generation in dyadic conversations," *arXiv:2412.04037*, 2024.

[70] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman, "Make-A-Video: Text-to-video generation without text-video data," in *The Eleventh International Conference on Learning Representations*, 2023.

[71] E. Molad, E. Horwitz, D. Valevski, A. R. Acha, Y. Matias, Y. Pritch, Y. Leviathan, and Y. Hoshen, "Dreamix: Video diffusion models are general video editors," *arXiv: 2302.01329*, 2023.

[72] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, "Modelscope text-to-video technical report," *arXiv: 2308.06571*, 2023.

[73] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7623–7633.

[74] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang, Y. Guo, T. Wu, C. Si, Y. Jiang, C. Chen, C. C. Loy, B. Dai, D. Lin, Y. Qiao, and Z. Liu, "LaVie: High-quality video generation with cascaded latent diffusion models," *International Journal of Computer Vision*, vol. 133, no. 5, pp. 3059–3078, 2025.

[75] T. Wu, Y. Zhang, X. Wang, X. Zhou, G. Zheng, Z. Qi, Y. Shan, and X. Li, "CustomCrafter: Customized Video Generation with Preserving Motion and Concept Composition Abilities," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 8469–8477.

[76] Y. Yang, F.-Y. Sun, L. Weihs, E. VanderBilt, A. Herrasti, W. Han, J. Wu, N. Haber, R. Krishna, L. Liu, C. Callison-Burch, M. Yatskar, A. Kembhavi, and C. Clark, "Holodeck: Language guided generation of 3D embodied AI environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[77] D. Xu, Y. Jiang, P. Wang, Z. Fan, Y. Wang, and Z. Wang, "NeuralLift-360: Lifting an in-the-wild 2d photo to a 3D object with 360° views," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 4479–4489.

[78] L. Melas-Kyriazi, I. Laina, C. Rupprecht, and A. Vedaldi, "Realfusion 360° reconstruction of any object from a single image," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 8446–8455.

[79] G. Qian, J. Mai, A. Hamdi, J. Ren, A. Siarohin, B. Li, H.-Y. Lee, I. Skorokhodov, P. Wonka, S. Tulyakov, and B. Ghanem, "Magic123: One image to high-quality 3D object generation using both 2d and 3D diffusion priors," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[80] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng, "Dreamgaussian: Generative gaussian splatting for efficient 3D content creation," in *The Twelfth International Conference on Learning Representations*, 2024.

[81] J. Ren, L. Pan, J. Tang, C. Zhang, A. Cao, G. Zeng, and Z. Liu, "DreamGaussian4D: Generative 4d gaussian splatting," *arXiv preprint arXiv:2312.17142*, 2023.

[82] Y. Zhao, M. Dasari, and T. Guo, "CleAR: Robust context-guided generative lighting estimation for mobile augmented reality," *arXiv preprint arXiv:2411.02179*, 2024, Available at https://arxiv.org/abs/2411.02179.

[83] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, and G. Wang, "Text classification via large language models," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[84] Y. Shi, H. Ma, W. Zhong, Q. Tan, G. Mai, X. Li, T. Liu, and J. Huang, "ChatGraph: Interpretable text classification by converting chatgpt knowledge to graphs," in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2023, pp. 515–520.

[85] B. Hou, J. O'Connor, J. Andreas, S. Chang, and Y. Zhang, "PromptBoosting: Black-box text classification with ten forward passes," in *Proceedings of the 40th International Conference on Machine Learning*, 23–29 Jul 2023, vol. 202, pp. 13309–13324.

[86] W. Ai, J. Li, Z. Wang, Y. Wei, T. Meng, and K. Li, "Contrastive multi-graph learning with neighbor hierarchical sifting for semi-supervised text classification," *Expert Systems with Applications*, vol. 266, pp. 125952, 2025.

[87] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1743–1753, 2023.

[88] J. O. Krugmann and J. Hartmann, "Sentiment analysis in the age of generative AI," *Customer Needs and Solutions*, vol. 11, no. 3, 2024.

[89] J. Hartmann, M. Heitmann, C. Siebert, and C. Schamp, "More than a feeling: Accuracy and application of sentiment analysis," *International Journal of Research in Marketing*, vol. 40, no. 1, pp. 75–87, 2023.

[90] D. Metzler, Y. Tay, D. Bahri, and M. Najork, "Rethinking search: making domain experts out of dilettantes," *SIGIR Forum*, vol. 55, no. 1, jul 2021.

[91] B. Yan, Y. Jiang, J. Wu, D. Wang, P. Luo, Z. Yuan, and H. Lu, "Universal instance perception as object discovery and retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 15325–15336.

[92] D. Lu, S.-Y. Wang, N. Kumari, R. Agarwal, M. Tang, D. Bau, and J.-Y. Zhu, "Content-based search for deep generative models," in *SIGGRAPH Asia 2023 Conference Papers*, 2023.

[93] S. Rajput, N. Mehta, A. Singh, R. Hulikal Keshavan, T. Vu, et al., "Recommender systems with generative retrieval," in *Advances in Neural Information Processing Systems*, 2023, vol. 36, pp. 10299–10315.

[94] X. Li, Y. Zhou, and Z. Dou, "UniGen: A unified generative framework for retrieval and question answering with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 8688–8696.

[95] Y. Li, N. Yang, L. Wang, F. Wei, and W. Li, "Learning to rank in generative retrieval," in *AAAI 2024*, 2024.

[96] P. Jin, H. Li, Z. Cheng, K. Li, X. Ji, C. Liu, L. Yuan, and J. Chen, "DiffusionRet: Generative text-video retrieval with diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 2470–2481.

[97] E. King, H. Yu, S. Lee, and C. Julien, "Sasha: Creative goal-oriented reasoning in smart homes with large language models," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 1, mar 2024.

[98] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "Snr-aware low-light image enhancement," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17693–17703.

[99] J. Liang, Y. Fan, X. Xiang, R. Ranjan, E. Ilg, S. Green, J. Cao, K. Zhang, R. Timofte, and L. Gool, "Recurrent video restoration transformer with guided deformable attention," in *Advances in Neural Information Processing Systems*, 2022.

[100] T. Wang, K. Zhang, T. Shen, W. Luo, B. Stenger, and T. Lu, "Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 2654–2662.

[101] R. Lin, N. Anantrasirichai, A. Malyugina, and D. Bull, "A spatio-temporal aligned sunet model for low-light video enhancement," in *IEEE International Conference on Image Processing*, 2024.

[102] G. Youk, J. Oh, and M. Kim, "FMA-Net: Flow-guided dynamic filtering and iterative feature refinement with multi-attention for joint video super-resolution and deblurring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[103] J. Liang, J. Cao, Y. Fan, K. Zhang, R. Ranjan, Y. Li, R. Timofte, and L. Van Gool, "Vrt: A video restoration transformer," *IEEE Transactions on Image Processing*, vol. 33, pp. 2171–2182, 2024.

[104] J. HOU, Z. Zhu, J. Hou, H. LIU, H. Zeng, and H. Yuan, "Global structure-aware diffusion process for low-light image enhancement," in *Advances in Neural Information Processing Systems*, 2023, vol. 36, pp. 79734–79747.

[105] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 12302–12311.

[106] H. Jiang, A. Luo, H. Fan, S. Han, and S. Liu, "Low-light image enhancement with wavelet-based diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1–14, 2023.

[107] R. Lin, Q. Sun, and N. Anantrasirichai, "Low-light video enhancement with conditional diffusion models and wavelet interscale attentions," in *Proceedings of the 21st ACM SIGGRAPH Conference on Visual Media Production*, 2024, pp. 1–10.

[108] S. Yang, M. Ding, Y. Wu, Z. Li, and J. Zhang, "Implicit neural representation for cooperative low-light image enhancement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 12918–12927.

[109] Y. Deng, F. Tang, W. Dong, C. Ma, X. Pan, L. Wang, and C. Xu, "StyTr2: Image style transfer with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11326–11336.

[110] J. Moon, T. Moon, and W. Seo, "Generalizable style transfer for implicit neural representation," in *International Conference on Learning Representations (ICLR)*, 2023.

[111] J. Chung, S. Hyun, and J.-P. Heo, "Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 8795–8805.

[112] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu, "Inversion-based style transfer with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 10146–10156.

[113] W. Chai, X. Guo, G. Wang, and Y. Lu, "StableVideo: Text-driven consistency-aware diffusion video editing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 23040–23050.

[114] S. Kim, Y. Min, Y. Jung, and S. Kim, "Controllable style transfer via test-time training of implicit neural representation," *Pattern Recognition*, vol. 146, pp. 109988, 2024.

[115] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 1833–1844.

[116] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2022, pp. 457–466.

[117] C. Liu, H. Yang, J. Fu, and X. Qian, "Learning trajectory-aware transformer for video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5687–5696.

[118] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 22367–22377.

[119] Y. Li, Y. Fan, X. Xiang, R. R. Denis Demandolx, R. Timofte, , and L. V. Gool, "Efficient and explicit modelling of image hierarchies for image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[120] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park, "Scaling up gans for text-to-image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[121] Y. Xu, T. Park, R. Zhang, Y. Zhou, E. Shechtman, F. Liu, J.-B. Huang, and D. Liu, "VideoGigaGAN: Towards detail-rich video super-resolution," *arXiv:2404.12388*, 2024.

[122] J. Wang, Z. Lin, M. Wei, Y. Zhao, C. Yang, C. C. Loy, and L. Jiang, "SeedVR: Seeding infinity in diffusion transformer towards generic video restoration," in *CVPR*, 2025.

[123] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4713–4726, 2023.

[124] E. Moliner, J. Lehtinen, and V. Välimäki, "Solving audio inverse problems with a diffusion model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[125] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang, "Implicit diffusion models for continuous super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 10021–10030.

[126] C. Cao, H. Yue, X. Liu, and J. Yang, "Zero-shot video restoration and enhancement using pre-trained image diffusion model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 1935–1943.

[127] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8624–8634.

[128] B. Fei, Z. Lyu, L. Pan, J. Zhang, W. Yang, T. Luo, B. Zhang, and B. Dai, "Generative diffusion prior for unified image restoration and enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 9935–9946.

[129] Y. Yin, D. Xu, C. Tan, P. Liu, Y. Zhao, and Y. Wei, "CLE Diffusion: Controllable light enhancement diffusion model," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, p. 8145–8156.

[130] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17662–17672.

[131] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5718–5729.

[132] H. Yang, L. Pan, Y. Yang, R. Hartley, and M. Liu, "LDP: Language-driven dual-pixel image defocus deblurring network," *arXiv: 2307.09815*, 2023.

[133] C. Morris, N. Anantrasirichai, F. Zhang, and D. Bull, "DaBiT: Depth and blur informed transformer for video deblurring," in *IEEE/CVF Winter Conference on Applications of Computer Vision Workshop*, 2025.

[134] G. Yu, A. Li, H. Wang, Y. Wang, Y. Ke, and C. Zheng, "Dbt-net: Dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2629–2644, 2022.

[135] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *IEEE Transactions on Image Processing*, vol. 32, pp. 1927–1941, 2023.

[136] J. Xu, X. Hu, L. Zhu, Q. Dou, J. Dai, Y. Qiao, and P.-A. Heng, "Video dehazing via a multi-range temporal alignment network with physical prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 18053–18062.

[137] Z. Mao, A. Jaiswal, Z. Wang, and S. H. Chan, "Single frame atmospheric turbulence mitigation: A benchmark study and a new physics-inspired transformer model," in *ECCV*, 2022.

[138] X. Zhang, Z. Mao, N. Chimitt, and S. H. Chan, "Imaging through the atmosphere using turbulence mitigation transformer," *IEEE Transactions on Computational Imaging*, vol. 10, pp. 115–128, 2024.

[139] Z. Zou and N. Anantrasirichai, "DeTurb: Atmospheric turbulence mitigation with deformable 3d convolutions and 3d swin transformers," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2024.

[140] W. Fang, J. Fan, Y. Zheng, J. Weng, Y. Tai, and J. Li, "Guided real image dehazing using ycbcr color space," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 2906–2914.

[141] H. Yue, C. Cao, L. Liao, and J. Yang, "RViDeformer: Efficient raw video denoising transformer with a larger benchmark dataset," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

[142] Y. Jin, X. Ma, R. Zhang, H. Chen, Y. Gu, P. Ling, and E. Chen, "Masked video pretraining advances real-world video denoising," *IEEE Transactions on Multimedia*, vol. 27, pp. 622–636, 2025.

[143] Y. Shi, B. Xia, X. Jin, X. Wang, T. Zhao, X. Xia, X. Xiao, and W. Yang, "VmambaIR: Visual state space model for image restoration," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[144] C. Yang, L. Liang, and Z. Su, "Real-world denoising via diffusion model," *arXiv preprint arXiv:2305.04457*, 2023.

[145] N. G. Nair, K. Mei, and V. M. Patel, "AT-DDPM: Restoring faces degraded by atmospheric turbulence using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023, pp. 3434–3443.

[146] A. Jaiswal, X. Zhang, S. H. Chan, and Z. Wang, "Physics-driven turbulence image restoration with stochastic refinement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 12170–12181.

[147] H. Feng, H. Zhou, T. Ye, S. Chen, and L. Zhu, "Residual diffusion deblurring model for single image defocus deblurring," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 2960–2968.

[148] W. Jiang, V. Boominathan, and A. Veeraraghavan, "NeRT: Implicit neural representations for unsupervised atmospheric turbulence mitigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023, pp. 4236–4243.

[149] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "MAT: Mask-aware transformer for large hole image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10758–10768.

[150] Q. Liu, Z. Tan, D. Chen, Q. Chu, X. Dai, Y. Chen, M. Liu, L. Yuan, and N. Yu, "Reduce information loss in transformers for pluralistic image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11347–11357.

[151] J. Ren, Q. Zheng, Y. Zhao, X. Xu, and C. Li, "DLFormer: Discrete latent transformer for video inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 3511–3520.

[152] S. Zhou, C. Li, K. C. Chan, and C. C. Loy, "ProPainter: Improving propagation and transformer for video inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 10477–10486.

[153] W. Huang, Y. Deng, S. Hui, Y. Wu, S. Zhou, and J. Wang, "Sparse self-attention transformer for image inpainting," *Pattern Recognition*, vol. 145, pp. 109897, 2024.

[154] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.

[155] D. Rao, T. Xu, and X.-J. Wu, "TGFuse: An infrared and visible image fusion approach based on transformer and generative adversarial network," *IEEE Transactions on Image Processing*, pp. 1–1, 2023.

[156] J. Liu, Z. Liu, G. Wu, L. Ma, R. Liu, W. Zhong, Z. Luo, and X. Fan, "Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 8115–8124.

[157] H. Li and X.-J. Wu, "CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach," *Information Fusion*, vol. 103, pp. 102147, 2024.

[158] Z. Zhao, H. Bai, Y. Zhu, J. Zhang, S. Xu, Y. Zhang, K. Zhang, D. Meng, R. Timofte, and L. Van Gool, "DDFM: Denoising diffusion model for multi-modality image fusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 8082–8093.

[159] X. Shi, Z. Huang, F.-Y. Wang, W. Bian, D. Li, Y. Zhang, M. Zhang, K. C. Cheung, S. See, H. Qin, J. Dai, and H. Li, "Motion-I2V: Consistent and controllable image-to-video generation with explicit motion modeling," in *ACM SIGGRAPH Conference Papers*, 2024.

[160] J. Guo, D. Zhang, X. Liu, Z. Zhong, Y. Zhang, P. Wan, and D. Zhang, "LivePortrait: Efficient portrait animation with stitching and retargeting control," *arXiv preprint arXiv:2407.03168*, 2024.

[161] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[162] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4015–4026.

[163] L. Ke, M. Ye, M. Danelljan, Y. liu, Y.-W. Tai, C.-K. Tang, and F. Yu, "Segment anything in high quality," in *Advances in Neural Information Processing Systems*, 2023, vol. 36, pp. 29914–29934.

[164] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, "SegGPT: Towards segmenting everything in context," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1130–1140.

[165] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. J. Lee, "Segment everything everywhere all at once," in *Advances in Neural Information Processing Systems*. 2023, vol. 36, pp. 19769–19782, Curran Associates, Inc.

[166] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024.

[167] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[168] T. Zhang, X. Tian, Y. Zhou, S. Ji, X. Wang, X. Tao, Y. Zhang, P. Wan, Z. Wang, and Y. Wu, "DVIS++: Improved decoupled framework for universal video segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[169] W. Wu, Y. Zhao, M. Z. Shou, H. Zhou, and C. Shen, "DiffuMask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 1206–1217.

[170] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 2955–2966.

[171] Z. Gu, H. Chen, and Z. Xu, "Diffusioninst: Diffusion model for instance segmentation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 2730–2734.

[172] R. Gong, Q. Wang, M. Danelljan, D. Dai, and L. Van Gool, "Continuous pseudo-label rectified domain adaptive semantic segmentation with implicit neural representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 7225–7235.

[173] J. Cen, Z. Zhou, J. Fang, c. yang, W. Shen, L. Xie, D. Jiang, X. ZHANG, and Q. Tian, "Segment anything in 3D with NeRFs," in *Advances in Neural Information Processing Systems*, 2023, vol. 36, pp. 25971–25990.

[174] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[175] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations (ICLR)*, 2021.

[176] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 3156–3165.

[177] K.-C. Huang, T.-H. Wu, H.-T. Su, and W. H. Hsu, "MonoDTR: Monocular 3D object detection with depth-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[178] L. Zhao, N. B. Gundavarapu, L. Yuan, H. Zhou, S. Yan, J. J., et al., "VideoPrism: A foundational visual encoder for video understanding," in *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[179] E. Im, C. Jee, and J. K. Lee, "GATE3D: Generalized Attention-based Task-synergized Estimation in 3D," *arXiv preprint arXiv:2504.11014*, 2025.

[180] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025.

[181] A. C. Li, M. Prabhudesai, S. Duggal, E. Brown, and D. Pathak, "Your diffusion model is secretly a zero-shot classifier," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 2206–2217.

[182] S. Chen, P. Sun, Y. Song, and P. Luo, "Diffusiondet: Diffusion model for object detection," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 19773–19786.

[183] H. Zhang, Z. Wang, D. Zeng, Z. Wu, and Y.-G. Jiang, "DiffusionAD: Norm-guided one-step denoising diffusion for anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2025.

[184] Z. Wu, L. Zhu, Z. Yin, X. Xu, J. Zhu, X. Wei, and X. Yang, "MAFCD: Multi-level and adaptive conditional diffusion model for anomaly detection," *Information Fusion*, vol. 118, pp. 102965, 2025.

[185] T. Meinhardt, A. Kirillov, L. Leal-Taixé, and C. Feichtenhofer, "Trackformer: Multi-object tracking with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 8844–8854.

[186] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "Motr: End-to-end multiple-object tracking with transformer," in *European Conference on Computer Vision (ECCV)*, 2022.

[187] Y. Cui, C. Jiang, L. Wang, and G. Wu, "Mixformer: End-to-end tracking with iterative mixed attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13608–13618.

[188] C. Mayer, M. Danelljan, G. Bhat, M. Paul, D. P. Paudel, F. Yu, and L. Van Gool, "Transforming model prediction for tracking," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 8721–8730.

[189] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, "Track anything: Segment anything meets videos," *arXiv:2304.11968*, 2023.

[190] X. Chen, H. Peng, D. Wang, H. Lu, and H. Hu, "Seqtrack: Sequence to sequence learning for visual object tracking," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14572–14581.

[191] Y. Zhang, T. Wang, and X. Zhang, "MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22056–22065.

[192] A. Yi and N. Anantrasirichai, "A comprehensive study of object tracking in low-light environments," *Sensors*, vol. 24, no. 14, 2024.

[193] B. Kang, X. Chen, S. Lai, Y. Liu, Y. Liu, and D. Wang, "Exploring enhanced contextual information for video-level object tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[194] R. Luo, Z. Song, L. Ma, J. Wei, W. Yang, and M. Yang, "DiffusionTrack: Diffusion model for multi-object tracking," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, pp. 3991–3999, Mar. 2024.

[195] F. Xie, Z. Wang, and C. Ma, "DiffusionTrack: Point set diffusion model for visual object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 19113–19124.

[196] R. Zhang, D. Cai, L. Qian, Y. Du, H. Lu, and Y. Zhang, "DiffusionTracker: Targets denoising based on diffusion model for visual tracking," in *Pattern Recognition and Computer Vision*, 2024, pp. 225–237.

[197] H. Jung, Z. Hui, L. Luo, H. Yang, F. Liu, S. Yoo, R. Ranjan, and D. Demandolx, "AnyFlow: Arbitrary scale optical flow with implicit neural representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 5455–5465.

[198] D. Wang, X. Cui, X. Chen, Z. Zou, T. Shi, S. Salcudean, Z. J. Wang, and R. Ward, "Multi-view 3D reconstruction with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 5722–5731.

[199] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "Lite-Mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 18537–18546.

[200] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[201] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[202] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," in *Advances in Neural Information Processing Systems*, 2024.

[203] D. Liu, Z. Wang, and P. Chen, "DSEM-NeRF: Multimodal feature fusion and global–local attention for enhanced 3d scene reconstruction," *Information Fusion*, vol. 115, pp. 102752, 2025.

[204] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-NeRF 360: Unbounded anti-aliased neural radiance fields," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5460–5469.

[205] Y. Ji, Z. Chen, E. Xie, L. Hong, X. Liu, Z. Liu, T. Lu, Z. Li, and P. Luo, "DDP: Diffusion model for dense visual prediction," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 21684–21695.

[206] J. Wynn and D. Turmukhambetov, "DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[207] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 9492–9502.

[208] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-NeRF: Neural radiance fields for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[209] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, July 2022.

[210] B. Mildenhall, P. Hedman, R. Martin-Brualla, P. P. Srinivasan, and J. 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 (CVPR)*, June 2022, pp. 16190–16199.

[211] J. Fang, T. Yi, X. Wang, L. Xie, X. Zhang, W. Liu, M. Nießner, and Q. Tian, "Fast dynamic radiance fields with time-aware neural voxels," in *SIGGRAPH Asia 2022 Conference Papers*, 2022.

[212] H. Guo, S. Peng, H. Lin, Q. Wang, G. Zhang, H. Bao, and X. Zhou, "Neural 3D scene reconstruction with the manhattan-world assumption," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5511–5520.

[213] T. Hu, S. Liu, Y. Chen, T. Shen, and J. Jia, "EfficientNeRF - Efficient neural radiance fields," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12892–12901.

[214] J.-W. Liu, Y.-P. Cao, J. Z. Wu, W. Mao, Y. Gu, R. Zhao, J. Keppo, Y. Shan, and M. Z. Shou, "Dynvideo-e: Harnessing dynamic nerf for large-scale motion- and view-change human-centric video editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 7664–7674.

[215] A. Azzarelli, N. Anantrasirichai, and D. R. Bull, "WavePlanes: A compact wavelet representation for dynamic neural radiance fields," *arXiv preprint arXiv:2312.02218*, 2023.

[216] Y. Zhan, Z. Li, M. Niu, Z. Zhong, S. Nobuhara, K. Nishino, and Y. Zheng, "KFD-NeRF: Rethinking dynamic nerf with kalman filter," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[217] X. Tang, M. Yang, P. Sun, H. Li, Y. Dai, F. Zhu, and H. Lee, "Parenerf: Toward fast large-scale dynamic nerf with patch-based reference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 5428–5438.

[218] Sara Fridovich-Keil and Giacomo Meanti, F. R. Warburg, B. Recht, and A. Kanazawa, "K-planes: Explicit radiance fields in space, time, and appearance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[219] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, July 2023.

[220] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and W. Xinggang, "4d gaussian splatting for real-time dynamic scene rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[221] H. Yu, J. Julin, Z. A. Milacski, K. Niinuma, and L. A. Jeni, "CoGS: Controllable gaussian splatting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 21624–21633.

[222] Y.-H. Huang, Y.-T. Sun, Z. Yang, X. Lyu, Y.-P. Cao, and X. Qi, "SC-GS: Sparse-controlled gaussian splatting for editable dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 4220–4230.

[223] H. Wang, N. Anantrasirichai, F. Zhang, and D. Bull, "UW-GS: Distractor-aware 3d gaussian splatting for enhanced underwater scene reconstruction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[224] H. Junkawitsch, G. Sun, H. Zhu, C. Theobalt, and M. Habermann, "EVA: Expressive Virtual Avatars from Multi-view Videos," in *ACM SIGGRAPH 2025 Conference Proceedings*. 2025, ACM.

[225] H. Kong, X. Yang, and X. Wang, "Efficient gaussian splatting for monocular dynamic scene rendering via sparse time-variant attribute modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 4374–4382.

[226] Y. Zhu, Y. Yang, and T. Cohen, "Transformer-based transform coding," in *International Conference on Learning Representations*, 2022.

[227] R. Zou, C. Song, and Z. Zhang, "The devil is in the details: Window-based attention for image compression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17492–17501.

[228] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-cnn architectures," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14388–14397.

[229] M. Careil, M. J. Muckley, J. Verbeek, and S. Lathuilière, "Towards image compression with perfect realism at ultra-low bitrates," in *The Twelfth International Conference on Learning Representations*, 2023.

[230] R. Yang and S. Mandt, "Lossy image compression with conditional diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[231] E. Hoogeboom, E. Agustsson, F. Mentzer, L. Versari, G. Toderici, and L. Theis, "High-fidelity image compression with score-based generative models," *arXiv preprint arXiv:2305.18231*, 2023.

[232] N. F. Ghouse, J. Petersen, A. Wiggers, T. Xu, and G. Sautiere, "A residual diffusion model for high perceptual quality codec augmentation," *arXiv preprint arXiv:2301.05489*, 2023.

[233] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in neural information processing systems*, vol. 33, pp. 7462–7473, 2020.

[234] E. Dupont, A. Golinski, M. Alizadeh, Y. W. Teh, and A. Doucet, "COIN: Compression with implicit neural representations," in *ICLR Workshop in Neural Compression*, 2021.

[235] E. Dupont, H. Loya, M. Alizadeh, A. Golinski, Y. W. Teh, and A. Doucet, "Coin++: Neural compression across modalities," *Transactions on Machine Learning Research*, 2022.

[236] Y. Strümpler, J. Postels, R. Yang, L. V. Gool, and F. Tombari, "Implicit neural representations for image compression," in *European Conference on Computer Vision*. Springer, 2022, pp. 74–91.

[237] J. Xiang, K. Tian, and J. Zhang, "Mimt: Masked image modeling transformer for video compression," in *International Conference on Learning Representations*, 2022.

[238] F. Mentzer, G. D. Toderici, D. Minnen, S. Caelles, S. J. Hwang, M. Lucic, and E. Agustsson, "Vct: A video compression transformer," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13091–13103, 2022.

[239] B. Li, Y. Liu, X. Niu, B. Bai, L. Deng, and D. Gündüz, "Extreme video compression with pre-trained diffusion models," *arXiv preprint arXiv:2402.08934*, 2024.

[240] H. Chen, B. He, H. Wang, Y. Ren, S. N. Lim, and A. Shrivastava, "NeRV: Neural representations for videos," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21557–21568, 2021.

[241] Y. Bai, C. Dong, C. Wang, and C. Yuan, "PS-NeRV: Patch-wise stylized neural representations for videos," in *IEEE International Conference on Image Processing*. IEEE, 2023, pp. 41–45.

[242] H. Kim, M. Bauer, L. Theis, J. R. Schwarz, and E. Dupont, "C3: High-performance and low-complexity neural compression from a single image or video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9347–9358.

[243] T. Leguay, T. Ladune, P. Philippe, and O. Déforges, "Cool-chic video: Learned video coding with 800 parameters," in *2024 Data Compression Conference (DCC)*, 2024, pp. 23–32.

[244] H. M. Kwan, G. Gao, F. Zhang, A. Gower, and D. Bull, "NVRC: Neural video representation compression," in *Advances in Neural Information Processing Systems*, 2024.

[245] G. Gao, H. M. Kwan, F. Zhang, and D. Bull, "PNVC: Towards practical INR-based video compression," *arXiv preprint arXiv:2409.00953*, 2024.

[246] H. Ruan, Y. Shao, Q. Yang, L. Zhao, and D. Niyato, "Point cloud compression with implicit neural representations: A unified framework," *arXiv preprint arXiv:2405.11493*, 2024.

[247] H. M. Kwan, F. Zhang, A. Gower, and D. Bull, "Immersive video compression using implicit neural representations," in *Picture Coding Symposium*, 2024.

[248] M. Cheon, S.-J. Yoon, B. Kang, and J. Lee, "Perceptual image quality assessment with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 433–442.

[249] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 1220–1230.

[250] J. Shi, P. Gao, and J. Qin, "Transformer-based no-reference image quality assessment via supervised contrastive learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 4829–4837.

[251] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin, "Fast-VQA: Efficient end-to-end video quality assessment with fragment sampling," in *European conference on computer vision*. Springer, 2022, pp. 538–554.

[252] C. Feng, D. Danier, F. Zhang, and D. Bull, "Rankdvqa: Deep vqa based on ranking-inspired hybrid training," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1648–1658.

[253] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20144–20154.

[254] C. He, Q. Zheng, R. Zhu, X. Zeng, Y. Fan, and Z. Tu, "COVER: A comprehensive video quality evaluator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5799–5809.

[255] T. Peng, C. Feng, D. Danier, F. Zhang, B. Vallade, A. Mackin, and D. Bull, "RMT-BVQA: Recurrent memory transformer-based blind video quality assessment for enhanced video content," in *European Conference on Computer Vision (ECCV) Workshop on Advances in Image Manipulation*, 2024.

[256] S. C. Shapiro and D. Eckroth, Eds., *Encyclopedia of Artificial Intelligence*, vol. 1, John Wiley & Sons, New York, 1987.

[257] D. Ippolito, A. Yuan, A. Coenen, and S. Burnam, "Creative writing with an AI-powered writing assistant: Perspectives from professional writers," *arXiv:2211.05030*, 2022.

[258] A. Azzarelli, N. Anantrasirichai, and D. Bull, "Intelligent cinematography: a review of ai research for cinematographic production," *Artificial Intelligence Review*, vol. 58, no. 108, 2025.

[259] A. Guo, P. Pataranutaporn, and P. Maes, "Exploring the interaction of creative writers with AI-powered writing tools," *arXiv:2402.12814*, 2024.

[260] P. Mirowski, K. W. Mathewson, J. Pittman, and R. Evans, "Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[261] C. Beckett and M. Yaseen, "Generating change: A global survey of what news organisations are doing with AI," Tech. Rep., JournalismAI, 2023.

[262] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on deep learning-based image captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 539–559, 2023.

[263] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021.

[264] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2024.

[265] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 6309–6318.

[266] Z. Wang, Q. Xie, T. Li, H. Du, L. Xie, P. Zhu, and M. Bi, "One-shot voice conversion for style transfer based on speaker adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6792–6796.

[267] Y. Huang, J. Huang, Y. Liu, M. Yan, J. Lv, J. Liu, W. Xiong, H. Zhang, L. Cao, and S. Chen, "Diffusion model-based image editing: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 6, pp. 4409–4437, 2025.

[268] F. Xu, T. Zhou, T. Nguyen, H. Bao, C. Lin, and J. Du, "Integrating augmented reality and llm for enhanced cognitive support in critical audio communications," *International Journal of Human-Computer Studies*, vol. 194, pp. 103402, 2025.

[269] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre, E. VanderBilt, A. Kembhavi, C. Vondrick, G. Gkioxari, K. Ehsani, L. Schmidt, and A. Farhadi, "Objaverse-XL: A universe of 10m+ 3D objects," in *Advances in Neural Information Processing Systems (NeurIPS)*, 7 2023.

[270] S. Feizi, M. Hajiaghayi, K. Rezaei, and S. Shin, "Online advertisements with llms: Opportunities and challenges," *arXiv preprint arXiv:2311.07601*, 2023.

[271] J. Ryu, K. Kim, D. Heo, H. Song, C. Oh, and B. Suh, "Cinema multiverse lounge: Enhancing film appreciation via multi-agent conversations," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 2025, CHI '25, ACM.

[272] X. Li, J. Jin, Y. Zhou, Y. Zhang, P. Zhang, Y. Zhu, and Z. Dou, "From matching to generation: A survey on generative information retrieval," *ACM Trans. Inf. Syst.*, vol. 43, no. 3, May 2025.

[273] A. Y. Chua, A. Pal, and S. Banerjee, "Ai-enabled investment advice: Will users buy it?," *Computers in Human Behavior*, vol. 138, pp. 107481, 2023.

[274] A. Aggarwal, "Evolution of recommendation systems in the age of generative ai," *International Journal of Science and Research Archive*, vol. 14, no. 01, pp. 485–492, 2025.

[275] E. Brynjolfsson, D. Li, and L. Raymond, "Generative AI at work," Tech. Rep., National Bureau of Economic Research, April 2023, NBER Working Paper No. w31161. Available at SSRN: https://ssrn.com/abstract=4426942.

[276] M. Lee, K. I. Gero, J. J. Y. Chung, S. B. Shum, V. Raheja, H. Shen, S. Venugopalan, T. Wambsganss, D. Zhou, E. A. Alghamdi, T. August, A. Bhat, M. Z. Choksi, S. Dutta, J. L. Guo, M. N. Hoque, Y. Kim, S. Knight, S. P. Neshaei, A. Shibani, D. Shrivastava, L. Shroff, A. Sergeyuk, J. Stark, S. Sterman, S. Wang, A. Bosselut, D. Buschek, J. C. Chang, S. Chen, M. Kreminski, J. Park, R. Pea, E. H. R. Rho, Z. Shen, and P. Siangliulue, "A design space for intelligent and interactive writing assistants," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024.

[277] R. Sajja, Y. Sermet, M. Cikmaz, D. Cwiertny, and I. Demir, "Artificial intelligence-enabled intelligent assistant for personalized and adaptive learning in higher education," *Information*, vol. 15, no. 10, pp. 596, 2024.

[278] N. Jiang, S. Hasanzadeh, and V. G. Duffy, "Domain-tailored generative ai for personalized assistant," in *HCI International 2024 – Late Breaking Papers*, V. G. Duffy, Ed., vol. 15376 of *Lecture Notes in Computer Science*, pp. 227–237. 2025.

[279] S. Zhou, C. Li, and C. Change Loy, "Lednet: Joint low-light enhancement and deblurring in the dark," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., 2022, pp. 573–589.

[280] G. Huang, N. Anantrasirichai, F. Ye, Z. Qi, R. Lin, Q. Yang, and D. Bull, "Bayesian neural networks for one-to-many mapping in image enhancement," *arXiv preprint*, vol. arXiv:2501.14265, January 2025.

[281] N. Anantrasirichai, R. Lin, A. Malyugina, and D. Bull, "BVI-Lowlight: Fully registered benchmark dataset for low-light video enhancement," *arXiv preprint arXiv:2402.01970*, 2024.

[282] G. Huang, R. Lin, Y. Li, D. Bull, and N. Anantrasirichai, "Bvi-mamba: Video enhancement using a visual state-space model for low-light and underwater environments," in *Machine Learning from Challenging Data*. 2025, vol. 13460 of *Proceedings of SPIE*, pp. 74–81, SPIE.

[283] R. Lin, N. Anantrasirichai, G. Huang, J. Lin, Q. Sun, A. Malyugina, and D. Bull, "BVI-RLV: A fully registered dataset and benchmarks for low-light video enhancement," *arXiv preprint arXiv:2407.03535*, 2024.

[284] N. Safonov, A. Bryncev, A. Moskalenko, D. Kulikov, et al., "NTIRE 2025 challenge on UGC video enhancement: Methods and results," *arXiv preprint arXiv:2505.03007*, 2025.

[285] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "Basicvsr++: Improving video super-resolution with enhanced propagation and alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5972–5981.

[286] X. Zhou, Y. Zheng, and J. Yang, "Bridging the metrics gap in image style transfer: A comprehensive survey of models and criteria," *Neurocomputing*, vol. 624, pp. 129430, 2025.

[287] M. V. Conde, E. Zamfir, R. Timofte, D. Motilla, et al., "Efficient deep models for real-time 4k image super-resolution. ntire 2023 benchmark and report," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 1495–1521.

[288] B. Bilecen and M. Ayazoglu, "Bicubic++: Slim, slimmer, slimmest designing an industry-grade super-resolution network," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, jun 2023, pp. 1623–1632.

[289] Z. Chen, K. Liu, J. Gong, J. Wang, et al., "Ntire 2025 challenge on image super-resolution (×4): Methods and results," *arXiv preprint arXiv:2504.14582*, 2025.

[290] B. B. Moser, A. S. Shanbhag, F. Raue, S. Frolov, S. Palacio, and A. Dengel, "Diffusion models, image super-resolution, and everything: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2024.

[291] D. Fuoli, M. Danelljan, R. Timofte, and L. Van Gool, "Fast online video super-resolution with deformable attention pyramid," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023, pp. 1735–1744.

[292] C. He, Y. Shen, C. Fang, F. Xiao, L. Tang, Y. Zhang, W. Zuo, Z. Guo, and X. Li, "Diffusion models in low-level vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 6, pp. 4630–4651, 2025.

[293] J. Pan, B. Xu, J. Dong, J. Ge, and J. Tang, "Deep discriminative spatial and temporal network for efficient video deblurring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 22191–22200.

[294] M. Choi, H. Lee, and H.-e. Lee, "Exploring positional characteristics of dual-pixel data for camera autofocus," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 13112–13122.

[295] Y. Yang, L. Pan, L. Liu, and M. Liu, "K3DN: Disparity-aware kernel estimation for dual-pixel defocus deblurring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 13263–13272.

[296] L. Sun, H. Guo, B. Ren, et al., "The tenth ntire 2025 image denoising challenge report," *arXiv preprint arXiv:2504.12276*, 2025.

[297] N. Anantrasirichai, "Atmospheric turbulence removal with complex-valued convolutional neural network," *Pattern Recognition Letters*, vol. 171, pp. 69–75, 2023.

[298] P. Hill, Z. Liu, and N. Anantrasirichai, "MAMAT: 3D mamba-based atmospheric turbulence removal and its object detection capability," in *21st IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS)*, 2025.

[299] P. Hill, N. Anantrasirichai, A. Achim, and D. Bull, "Deep learning techniques for atmospheric turbulence removal: A review," *Artificial Intelligence Review*, vol. 58, no. 101, 2025.

[300] C. Zheng, T.-J. Cham, and J. Cai, "Pluralistic image completion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1438–1447.

[301] Z. Zhang, Z. Hu, W. Deng, C. Fan, T. Lv, and Y. Ding, "DINet: Deformation inpainting network for realistic face visually dubbing on high resolution video," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 3543–3551.

[302] W. Quan, J. Chen, Y. Liu, and et al., "Deep learning-based image and video inpainting: A survey," *International Journal of Computer Vision*, 2024.

[303] O. Elharrouss, R. Damseh, A. N. Belkacem, S. Al-Maadeed, A. Saleous, and S. Bourennane, "Transformer-based image and video inpainting: Current challenges and future directions," *Artificial Intelligence Review*, vol. 58, pp. 124, 2025.

[304] S. Karim, G. Tong, J. Li, A. Qadir, U. Farooq, and Y. Yu, "Current advances and future perspectives of image fusion: A comprehensive review," *Information Fusion*, vol. 90, pp. 185–217, 2023.

[305] X. Zhang and Y. Demiris, "Visible and infrared image fusion using deep learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10535–10554, 2023.

[306] R. Tous, "Lester: Rotoscope animation through video object segmentation and tracking," *Algorithms*, vol. 17, no. 8, 2024.

[307] D. Baranchuk, A. Voynov, I. Rubachev, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," in *International Conference on Learning Representations*, 2022.

[308] J. Lin, N. Anantrasirichai, and D. Bull, "Feature denoising for low-light instance segmentation using weighted non-local blocks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2025.

[309] R. Goel, D. Sirikonda, S. Saini, and P. J. Narayanan, "Interactive segmentation of radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 4201–4211.

[310] Y. He, H. Yu, X. Liu, Z. Yang, W. Sun, S. Anwar, and A. Mian, "Deep learning based 3d segmentation in computer vision: A survey," *Information Fusion*, vol. 115, pp. 102722, 2025.

[311] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[312] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

[313] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[314] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[315] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-time end-to-end object detection," in *Advances in Neural Information Processing Systems*, 2024.

[316] Y. Li, N. Miao, L. Ma, F. Shuang, and X. Huang, "Transformer for object detection: Review and benchmark," *Engineering Applications of Artificial Intelligence*, vol. 126, pp. 107021, 2023.

[317] H. Kheddar, "Transformers and large language models for efficient intrusion detection systems: A comprehensive survey," *Information Fusion*, vol. 124, pp. 103347, 2025.

[318] Z. Song, L. Liu, F. Jia, Y. Luo, C. Jia, G. Zhang, L. Yang, and L. Wang, "Robustness-aware 3d object detection in autonomous driving: A review and outlook," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 11, pp. 15407–15436, 2024.

[319] W. Wu, Y. Zhao, H. Chen, Y. Gu, R. Zhao, Y. He, H. Zhou, M. Z. Shou, and C. Shen, "DatasetDM: Synthesizing data with perception annotations using diffusion models," in *Advances in Neural Information Processing Systems*. 2023, vol. 36, pp. 54683–54695, Curran Associates, Inc.

[320] H. Fang, B. Han, S. Zhang, S. Zhou, C. Hu, and W.-M. Ye, "Data augmentation for object detection via controllable diffusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 1257–1266.

[321] J. Kugarajeevan, T. Kokul, A. Ramanan, and S. Fernando, "Transformers in single object tracking: An experimental survey," *IEEE Access*, vol. 11, pp. 80297–80326, 2023.

[322] H. K. Cheng and A. G. Schwing, "Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model," in *ECCV*, 2022, vol. 13688, pp. 640–658.

[323] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[324] A. Azzarelli, N. Anantrasirichai, and D. R. Bull, "Exploring dynamic novel view synthesis technologies for cinematography," *arXiv: 2412.17532*, 2024.

[325] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[326] Y. Jiang, C. Yu, T. Xie, X. Li, Y. Feng, H. Wang, M. Li, H. Lau, F. Gao, Y. Yang, and C. Jiang, "VR-GS: A physical dynamics-aware interactive gaussian splatting system in virtual reality," in *ACM SIGGRAPH 2024 Conference Papers*, 2024.

[327] B. Fei, J. Xu, R. Zhang, Q. Zhou, W. Yang, and Y. He, "3d gaussian splatting as new era: A survey," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–20, 2024.

[328] D. Bull and F. Zhang, *Intelligent image and video compression: communicating pictures*, Academic Press, 2021.

[329] J. Ballé, V. Laparra, and E. P. Simoncelli, "Density modeling of images using a generalized normalization transformation," in *International Conference on Learning Representations (ICLR)*, 2016.

[330] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.

[331] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7939–7948.

[332] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 221–231.

[333] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11913–11924, 2020.

[334] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Asilomar Conf. Signals Syst. Comput.*, 2003, pp. 1398–1402.

[335] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "The NETFLIX tech blog: Toward a practical perceptual video quality metric," http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html, note = [Online; accessed 2018-08-04],.

[336] " Challenge on Learned Image Compression," https://www.compression.cc/, Accessed: 2024-09-23.

[337] D. Li, Y. Bai, K. Wang, J. Jiang, and X. Liu, "Semantic ensemble loss and latent refinement for high-fidelity neural image compression," in *2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2024, pp. 1–5.

[338] J. Ascenso, E. Alshina, and T. Ebrahimi, "The JPEG AI standard: Providing efficient human and machine visual data consumption," *IEEE Multimedia*, vol. 30, no. 1, pp. 100–111, 2023.

[339] E. Alshina, J. Ascenso, S. Esenlik, A. Karabutov, Y. Wu, and T. Solovyev, "JPEG AI: Future Plans and Timeline v2," *JPEG AI ISO/IEC JTC 1/SC29/WG1 N1100634*, 2024.

[340] A. Karabutov, Y. Wu, E. Alshina, and J. Ascenso, "JPEG AI sw v4.x status," *JPEG AI ISO/IEC JTC 1/SC29/WG1 M101081*, 2024.

[341] T. Li, M. Xu, R. Tang, Y. Chen, and Q. Xing, "DeepQTMT: A deep learning approach for fast QTMT-based CU partition of intra-mode VVC," *IEEE Transactions on Image Processing*, vol. 30, pp. 5377–5390, 2021.

[342] D. Jin, J. Lei, B. Peng, W. Li, N. Ling, and Q. Huang, "Deep affine motion compensation network for inter prediction in VVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3923–3933, 2021.

[343] Z. Feng, C. Jung, H. Zhang, Y. Liu, and M. Li, "Low complexity in-loop filter for VVC based on convolution and transformer," *IEEE Access*, 2024.

[344] H. Zhang, C. Jung, D. Zou, and M. Li, "WCDANN: A lightweight CNN post-processing filter for VVC-based video compression," *IEEE Access*, 2023.

[345] Y. Wang, T. Isobe, X. Jia, X. Tao, H. Lu, and Y.-W. Tai, "Compression-aware video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2012–2021.

[346] Y. Li, J. Li, C. Lin, K. Zhang, L. Zhang, F. Galpin, T. Dumas, H. Wang, M. Coban, J. Ström, et al., "Designs and implementations in neural network-based video coding," *arXiv preprint arXiv:2309.05846*, 2023.

[347] F. Galpin, Y. Li, Y. Li, J. N. Shingala, L. Wang, and Z. Xie, "NNVC software development AhG14," *Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, doc. no. JVET-AG0014*, January 2024.

[348] U. Joshi, Y. Chen, I. Yoo, S. Li, F. Yang, and D. Mukherjee, "Switchable cnns for in-loop restoration and super-resolution for av2," in *Applications of Digital Image Processing XLVI*. SPIE, 2023, vol. 12674, pp. 121–130.

[349] B. Kathariya, Z. Li, and G. Van der Auwera, "Joint pixel and frequency feature learning and fusion via channel-wise transformer for high-efficiency learned in-loop filter in vvc," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[350] A. Chadha and Y. Andreopoulos, "Deep perceptual preprocessing for video coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14852–14861.

[351] H. Tan, G. Xiang, X. Xie, and H. Jia, "Joint frame-level and block-level rate-perception optimized preprocessing for video coding," in *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, 2024.

[352] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An end-to-end deep video compression framework," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11006–11015.

[353] J. Li, B. Li, and Y. Lu, "Neural video compression with feature modulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26099–26108.

[354] L. Qi, Z. Jia, J. Li, B. Li, H. Li, and Y. Lu, "Long-term temporal context gathering for neural video compression," in *European Conference on Computer Vision (ECCV)*, 2024.

[355] Z. Hu, G. Lu, and D. Xu, "FVC: A new framework towards deep video compression in feature space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1502–1511.

[356] J. Li, B. Li, and Y. Lu, "Deep contextual video compression," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18114–18125, 2021.

[357] M. Khani, V. Sivaraman, and M. Alizadeh, "Efficient video compression via content-adaptive super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4521–4530.

[358] S. Oh, H. Yang, and E. Park, "Parameter-efficient instance-adaptive neural video compression," in *Asian Conference on Computer Vision*, 2024.

[359] J. Li, B. Li, and Y. Lu, "Neural video compression with diverse contexts," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22616–22626.

[360] Y.-H. Ho, C.-P. Chang, P.-Y. Chen, A. Gnutti, and W.-H. Peng, "CANF-VC: Conditional augmented normalizing flows for video compression," in *European Conference on Computer Vision*. Springer, 2022, pp. 207–223.

[361] W. Guo-Hua, J. Li, B. Li, and Y. Lu, "Evc: Towards real-time neural image compression with mask decay," in *International Conference on Learning Representations*, 2023.

[362] T. Peng, G. Gao, H. Sun, F. Zhang, and D. Bull, "Accelerating learnt video codecs with gradient decay and layer-wise distillation," in *2024 Picture Coding Symposium (PCS)*. IEEE, 2024, pp. 1–5.

[363] J. Nawała, Y. Jiang, F. Zhang, X. Zhu, J. Sole, and D. Bull, "BVI-AOM: A new training dataset for deep video compression optimization," in *IEEE Visual Communications and Image Processing (VCIP))*, 2024.

[364] Z. Li, M. Wang, H. Pi, K. Xu, J. Mei, and Y. Liu, "E-NeRV: Expedite neural video representation with disentangled spatial-temporal context," in *European Conference on Computer Vision*. Springer, 2022, pp. 267–284.

[365] F. Bossen, J. Boyce, K. Suehring, X. Li, and V. Seregin, "VTM common test conditions and software reference configurations for SDR video," in *the JVET meeting*, 2023, number JVET-T2010.

[366] "ISCAS 2024 grand challenge on neural network-based video coding," https://iscasnnvcgc.github.io/, Accessed: 2025-05-21.

[367] Y. Zhao, W. He, C. Jia, Q. Wang, J. Li, Y. Li, C. Lin, K. Zhang, L. Zhang, and S. Ma, "A neural-network enhanced video coding framework beyond ECM," in *2024 Data Compression Conference (DCC)*. IEEE, 2024, pp. 605–605.

[368] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[369] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[370] H. Siuzdak, F. Grötschla, and L. A. Lanzendörfer, "SNAC: Multi-scale neural audio codec," in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[371] D. Yang, H. Guo, Y. Wang, R. Huang, X. Li, X. Tan, X. Wu, and H. M. Meng, "UniAudio 1.5: Large language model-driven audio codec is a few-shot audio task learner," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[372] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, J. Shi, S. Zhao, J. Bian, X. Wu, et al., "UniAudio: An audio foundation model toward universal audio generation," *arXiv preprint arXiv:2310.00704*, 2023.

[373] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, pp. 1–52, 2020.

[374] Q. Zheng, Y. Fan, L. Huang, T. Zhu, J. Liu, Z. Hao, S. Xing, C.-J. Chen, X. Min, A. C. Bovik, et al., "Video quality assessment: A comprehensive survey," *arXiv preprint arXiv:2412.04508*, 2024.

[375] Z. Zhang, Y. Zhou, C. Li, B. Zhao, X. Liu, and G. Zhai, "Quality assessment in the era of large models: A survey," *arXiv preprint arXiv:2409.00031*, 2024.

[376] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, April 2004.

[377] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003.* Ieee, 2003, vol. 2, pp. 1398–1402.

[378] A. Rehman, K. Zeng, and Z. Wang, "Display device-adapted video quality-of-experience assessment," in *Human vision and electronic imaging XX.* SPIE, 2015, vol. 9394, pp. 27–37.

[379] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE transactions on image processing*, vol. 16, no. 9, pp. 2284–2298, 2007.

[380] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of electronic imaging*, vol. 19, no. 1, pp. 011006–011006, 2010.

[381] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *IEEE ICIP*, Sep. 2011, pp. 2505–2508.

[382] Y.-F. Ou, Z. Ma, T. Liu, and Y. Wang, "Perceptual quality assessment of video considering both frame rate and quantization artifacts," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 286–298, 2010.

[383] K. Zhu, C. Li, V. Asari, and D. Saupe, "No-reference video quality assessment based on artifact measurement and statistical analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 4, pp. 533–546, 2014.

[384] F. Zhang and D. R. Bull, "A perception-based hybrid model for video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1017–1028, 2015.

[385] H. L. F. von Helmholtz, *Handbook of Physiological Optics*, Voss, Hamburg and Leipzig, Germany, 1st edition, 1896.

[386] D. Kelly, "Visual contrast sensitivity," *Optica Acta: International Journal of Optics*, vol. 24, no. 2, pp. 107–129, 1977.

[387] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[388] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1676–1684.

[389] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[390] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Image quality assessment using contrastive learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 4149–4161, 2022.

[391] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019.

[392] M. Xu, J. Chen, H. Wang, S. Liu, G. Li, and Z. Bai, "C3dvqa: Full-reference video quality assessment with 3d convolutional neural network," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 4447–4451.

[393] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 219–234.

[394] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[395] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, C. Li, W. Sun, Q. Yan, G. Zhai, et al., "Q-bench: A benchmark for general-purpose foundation models on low-level vision," in *The Twelfth International Conference on Learning Representations*, 2024.

[396] H. Wu, Z. Zhang, W. Zhang, C. Chen, L. Liao, C. Li, Y. Gao, A. Wang, E. Zhang, W. Sun, et al., "Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels," in *International Conference on Machine Learning*, 2024.

[397] Y. Chen, L. Liu, and C. Ding, "X-iqe: explainable image quality evaluation for text-to-image generation with visual large language models," *arXiv preprint arXiv:2305.10843*, 2023.

[398] H. Zhu, X. Sui, B. Chen, X. Liu, Y. Fang, and S. Wang, "2AFC prompting of large multimodal models for image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[399] T. Miyata, "Zen-iqa: Zero-shot explainable and no-reference image quality assessment with vision language model," *IEEE Access*, vol. 12, pp. 70973–70983, 2024.

[400] W. Pan, Z. Yang, D. Liu, C. Fang, Y. Zhang, and P. Dai, "Quality-aware clip for blind image quality assessment," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2023, pp. 396–408.

[401] Z. Chen, H. Qin, J. Wang, C. Yuan, B. Li, W. Hu, and L. Wang, "Promptiqa: Boosting the performance and generalization for no-reference image quality assessment via prompts," in *European Conference on Computer Vision*. Springer, 2025, pp. 247–264.

[402] H. Wu, L. Liao, J. Hou, C. Chen, E. Zhang, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring opinion-unaware video quality assessment with semantic affinity criterion," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 366–371.

[403] H. Wu, L. Liao, A. Wang, C. Chen, J. Hou, W. Sun, Q. Yan, and W. Lin, "Towards robust text-prompted semantic criterion for in-the-wild video quality assessment," *arXiv preprint arXiv:2304.14672*, 2023.

[404] H. R. Sheikh, M. F. Sabir, , and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on image processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[405] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Color image database tid2013: Peculiarities and preliminary results," in *European Workshop on Visual Information Processing (EUVIP)*, 2013, pp. 106–111.

[406] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[407] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The konstanz natural video database (konvid-1k)," in *2017 Ninth International Conf. on Quality of Multimedia Experience (QoMEX)*. IEEE, 2017, pp. 1–6.

[408] Y. Wang, S. Inguva, and B. Adsumilli, "Youtube UGC dataset for video compression research," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–5.

[409] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. on Image Processing*, vol. 28, no. 2, pp. 612–627, 2018.

[410] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," *IEEE Access*, vol. 9, pp. 108069–108082, 2021.

[411] F. Zhou, W. Sheng, Z. Lu, and G. Qiu, "A database and model for the visual quality assessment of super-resolution videos," *IEEE Transactions on Broadcasting*, vol. 70, no. 2, pp. 516–532, 2024.

[412] Z. Chen, W. Sun, J. Jia, F. Lu, Z. Zhang, J. Liu, R. Huang, X. Min, and G. Zhai, "Band-2k: Banding artifact noticeable database for banding detection and quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6347–6362, 2024.

[413] C. Feng, D. Danier, F. Zhang, A. Mackin, A. Collins, and D. Bull, "Bvi-artefact: An artefact detection benchmark dataset for streamed videos," in *2024 Picture Coding Symposium (PCS)*. IEEE, 2024, pp. 1–5.

[414] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1040–1049.

[415] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, 2021.

[416] Q. Hou, A. Ghildyal, and F. Liu, "A perceptual quality metric for video frame interpolation," in *European Conf. on Computer Vision*. Springer, 2022, pp. 234–253.

[417] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Conviqt: Contrastive video quality estimator," *IEEE Transactions on Image Processing*, 2023.

[418] K. Zhao, K. Yuan, M. Sun, M. Li, and X. Wen, "Quality-aware pre-trained models for blind image quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22302–22313.

[419] L. Zhong, Z. Wang, and J. Shang, "LDB: A large language model debugger via verifying runtime execution step-by-step," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics: Findings.*, 2024.

[420] DeepSeek-AI, A. Liu, B. Feng, B. Xue, et al., "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[421] Q. Team, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[422] C. I. Council, "How createch added 1+1 to make £981m," 2021, Accessed: 2025-01-10.