Scene Text Synthesis and Visual Text Generation: A Survey

www.surveyx.cn

Abstract

Scene text synthesis and visual text generation represent a dynamic intersection of computer vision and natural language processing (NLP), facilitated by advancements in deep learning, generative adversarial networks (GANs), and optical character recognition (OCR). These technologies enable the generation and interpretation of text within images, enhancing applications in augmented reality, autonomous driving, and digital communication. Recent progress includes the development of sophisticated frameworks like SceneVTG and AnyText, which leverage multimodal large language models to ensure semantic coherence and visual realism. However, challenges persist in achieving model generalization and robustness, particularly due to the scarcity of high-quality, diverse datasets. The integration of advanced architectures and data augmentation strategies using large language models offers promising solutions to these challenges. Future research directions include enhancing model adaptability, optimizing computational efficiency, and expanding multilingual support to improve the fidelity and applicability of generated content. These advancements hold the potential to significantly enhance the capabilities of scene text synthesis and visual text generation, driving innovation and expanding their applicability across various domains. Continued exploration in this field is expected to yield more sophisticated models capable of handling complex visual and linguistic data, ultimately broadening the scope and impact of these technologies.

1 Introduction

1.1 Overview of Scene Text Synthesis and Visual Text Generation

Scene text synthesis and visual text generation are critical intersections of computer vision and natural language processing (NLP), focused on generating and interpreting text within visual contexts. These domains address the shortage of annotated datasets required for training complex neural networks by creating synthetic data that enhances model robustness. The generation of text images is essential for improving scene text recognition (STR) models, which often struggle with dataset variability [1].

In practical applications, scene text synthesis is crucial in areas such as autonomous driving and augmented reality, where accurate text recognition is vital for operational efficiency [2]. The ability to generate and manipulate text in images across diverse styles and languages represents significant progress in the field [3]. The evolution of deep learning architectures, particularly transformer networks and attention mechanisms, has accelerated advancements by optimizing image representation and integrating multimodal data.

Generative adversarial networks (GANs) have significantly contributed to visual text generation, achieving high visual realism and semantic coherence between textual descriptions and visual content [4]. However, challenges remain, particularly in synthesizing coherent motion and background features in video generation [5]. The integration of large pre-trained language models (PLMs) into

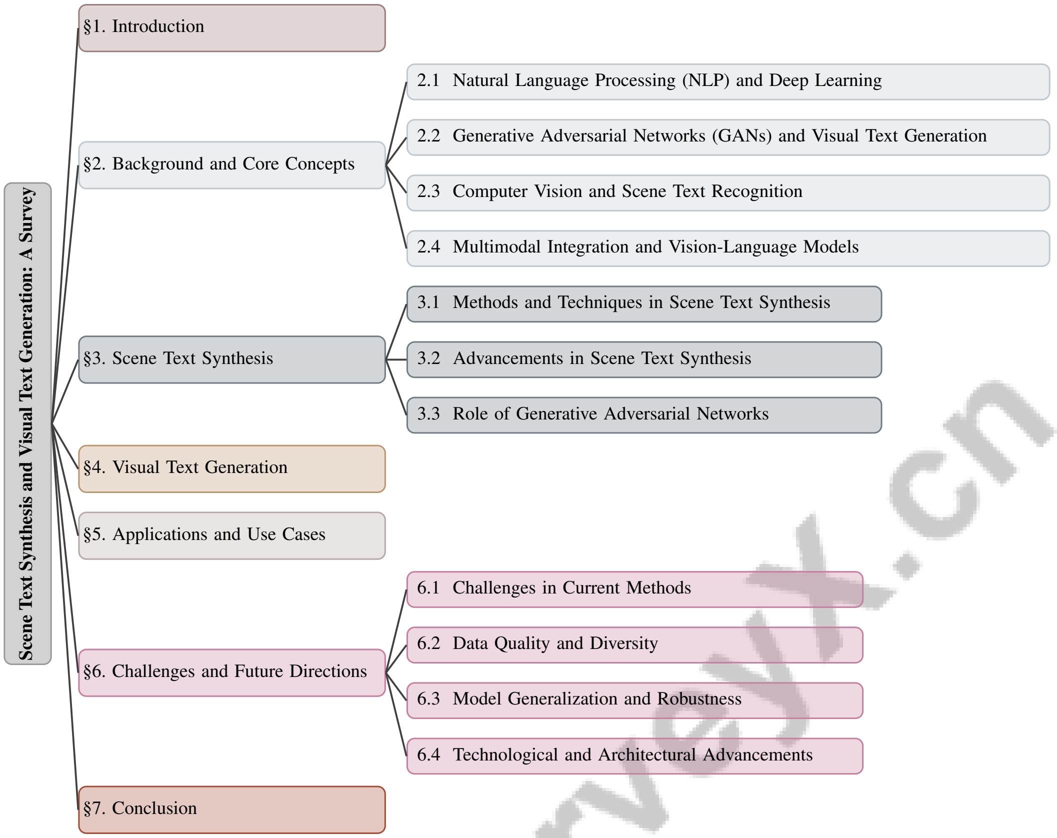

Scene Text Synthesis and Visual Text Generation: A Survey
§1. Introduction
§2. Background and Core Concepts
2.1 Natural Language Processing (NLP) and Deep Learning
2.2 Generative Adversarial Networks (GANs) and Visual Text Generation
2.3 Computer Vision and Scene Text Recognition
2.4 Multimodal Integration and Vision-Language Models
§3. Scene Text Synthesis
3.1 Methods and Techniques in Scene Text Synthesis
3.2 Advancements in Scene Text Synthesis
3.3 Role of Generative Adversarial Networks
§4. Visual Text Generation
§5. Applications and Use Cases
§6. Challenges and Future Directions
6.1 Challenges in Current Methods
6.2 Data Quality and Diversity
6.3 Model Generalization and Robustness
6.4 Technological and Architectural Advancements
§7. Conclusion

Figure 1: chapter structure

text generation processes helps bridge knowledge gaps and unify various methodologies to advance the field [6].

These technological advancements are set to enhance digital communication and support assistive technologies for the visually impaired, highlighting their importance in the future of human-computer interaction [7]. The development of adaptable text removal methods for various fonts, scripts, and languages underscores the need for flexible approaches in text manipulation within images [8]. Moreover, rapid progress in text-to-image generative models has transformed computer vision, enabling the synthesis of high-quality images from textual descriptions [9]. The challenge of visually translating scene text from a source language to a target language while maintaining visual features such as font and background exemplifies the practical applications and challenges in this domain [10].

1.2 Interconnected Technologies

The convergence of deep learning, generative adversarial networks (GANs), and optical character recognition (OCR) is crucial for advancing scene text synthesis and visual text generation. Deep learning architectures, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have significantly enhanced NLP, improving tasks like sentiment analysis and machine translation [11]. These models tackle challenges such as language ambiguity and contextual understanding by leveraging knowledge from various domains [12]. The pre-train and fine-tune paradigm, along with prompt-based learning, emphasizes the synergy between deep learning and NLP in text generation tasks [13]. This synergy is further enhanced by data augmentation techniques that utilize large language models (LLMs) to refine data and learning strategies [2].

GANs have been instrumental in visual text generation, improving the realism and semantic coherence of synthetic data [14]. Nevertheless, traditional GAN methods encounter difficulties with complex scenes, prompting the integration of variational autoencoders to impose semantic structures on

AI-generated, for reference only.

generated content. Hybrid frameworks that combine GANs with other models show promise in extracting both static and dynamic information, particularly for video generation tasks [15].

OCR complements these efforts by converting text within images into machine-readable formats, essential for retrieving information from visual data. However, OCR faces challenges in multilingual environments, often focusing predominantly on English text and treating localization and recognition as separate tasks [16]. This limitation underscores the need for unified multimodal models that can process both text and images [7]. The development of visual foundation models (VFMs), utilizing 3D transformer frameworks, exemplifies the interconnectedness of these technologies in managing both images and videos [17].

The interplay between deep learning, GANs, and OCR is vital for overcoming challenges in scene text synthesis and visual text generation. By harnessing these interconnected technologies, researchers can create robust models capable of handling diverse datasets and generating high-quality synthetic data, thus advancing the fields of computer vision and NLP [18]. The unification of generative and discriminative VFMs encompasses the evolution, methodologies, and applications of these models, including tasks such as text-to-image synthesis and image segmentation [15]. The integration of cutting-edge techniques in NLP and computer vision, as demonstrated in frameworks like PCIG, further illustrates the potential of these technologies to enhance the fidelity and coherence of generated content [9].

1.3 Structure of the Survey

This survey is structured to provide an in-depth examination of the interconnected fields of scene text synthesis and visual text generation, emphasizing their significance within computer vision and NLP. It begins with an introduction that highlights the importance of these fields and the role of technologies such as deep learning and GANs. The survey explores foundational concepts and key advancements in AI-driven text-to-image and text-to-video generation technologies, detailing essential deep learning and NLP techniques, including data preprocessing, neural network architectures, and evaluation metrics. It also addresses the challenges these technologies encounter and suggests potential future research directions, underscoring their promising applications across domains like video production, content creation, and digital marketing [19, 20, 13, 21, 22].

Subsequent sections delve into specific areas of interest. Section 3 focuses on scene text synthesis, discussing various methods, recent advancements, and the role of GANs in enhancing synthesis quality. Section 4 shifts to visual text generation, examining approaches and frameworks, the importance of coherence and contextual alignment, and innovations in text representation.

In Section 5, we highlight a range of practical applications and use cases, demonstrating their significance in fields such as augmented reality, which enhances user experiences by overlaying digital content onto the real world; assistive technologies that improve accessibility for individuals with disabilities; autonomous driving, where AI systems interpret and respond to driving environments; and AI-generated art, which utilizes advanced algorithms to create unique visual works from textual descriptions. These applications illustrate the transformative potential of AI technologies in reshaping industries and enhancing creative processes [22, 20, 23]. Section 6 addresses the challenges faced in current methods, emphasizing data quality, model generalization, and potential technological advancements. The survey concludes with a summary of key findings and suggestions for future research directions.The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Natural Language Processing (NLP) and Deep Learning

NLP and deep learning are pivotal in modern text synthesis and generation, revolutionizing the understanding and production of human language. The shift from traditional statistical methods to deep neural networks has expanded NLP applications, enhancing tasks like machine translation, sentiment analysis, and dialogue systems [24]. Integrating vision and language has shown potential in joint representation, boosting task performance across domains [25].

Convolutional Neural Networks (CNNs) are particularly effective in capturing intricate textual patterns, improving NLP task outcomes. However, generating coherent and contextually relevant sentences remains challenging due to the discrete nature of text inputs and the limitations of existing generative models in recognizing language's hierarchical structure [26]. The Prompt-Consistency Image Generation (PCIG) framework addresses this by combining large language models, knowledge graphs, and controllable diffusion models to ensure image generation consistency [9].

The pre-train and fine-tune paradigm exemplifies deep learning's integration into NLP, with models trained on vast unlabeled corpora and fine-tuned for specific tasks [24]. This approach, alongside prompt-based learning, efficiently utilizes pre-trained language models (PLMs) by framing NLP tasks as text generation problems. Models like GIT, which connect image encoders to text decoders, illustrate the seamless integration of visual and textual data for tasks such as caption generation [15].

Despite these advancements, challenges persist in enabling PLMs to continuously learn and adapt post-deployment, crucial for maintaining relevance and accuracy. Effective strategies for using PLMs to produce coherent, fluent, and contextually appropriate text from diverse data inputs are essential for further progress [24]. The synergy between NLP and deep learning continues to drive AI innovation, enhancing text synthesis and generation fidelity.

2.2 Generative Adversarial Networks (GANs) and Visual Text Generation

GANs have revolutionized visual text generation, enabling the creation of images that are both visually compelling and semantically aligned with their textual descriptions. The GAN architecture, with its generator-discriminator dynamic, enhances the realism and quality of outputs through adversarial training [14]. Frameworks like Semantics Disentangling GAN (SD-GAN) refine this by distilling common semantics while preserving detail diversity [27].

Innovative methodologies such as GlyphControl improve text-to-image model performance by introducing glyph conditional information, enhancing visual text generation fidelity [28]. These approaches leverage high-level sentence embeddings to guide the generative process, exemplified by adversarial and autoregressive model combinations [26].

Attention mechanisms and sequence-to-sequence models within GAN frameworks enhance the emotional and contextual relevance of generated content, as seen in multimodal applications like meme generation [29]. Diffusion-based frameworks like PCIG tackle alignment challenges between generated images and textual descriptions, ensuring consistency and reducing hallucinations [9].

Challenges remain, notably in text accuracy and legibility within generated images, with issues like misspellings due to tokenization limitations and insufficient learning in cross-attention modules. Ongoing refinement of GAN architectures is necessary to enhance their applicability in visual text generation [24]. Integrating OCR with NLP for post-processing further underscores the potential for improving data accuracy and extraction [16].

GANs play a crucial role in advancing visual text generation by merging textual and visual data to produce high-quality, semantically rich images. Deep learning advancements significantly enhance computer vision and NLP capabilities, establishing a robust framework for innovative multimodal data synthesis applications, including text-to-image and text-to-video generation, image captioning, and meme creation [6, 23, 20, 29, 22].

2.3 Computer Vision and Scene Text Recognition

Scene Text Recognition (STR) is integral to computer vision, focusing on deciphering text in natural environments. Challenges arise from diverse text appearances, layouts, and environmental factors like lighting and occlusion, which impact recognition accuracy [30]. Traditional STR methods often face inconsistencies in evaluation settings, leading to performance variations across models [31].

Recent STR advancements introduce innovative frameworks to address these challenges. The WeText framework uses a regression-based deep network to enhance character detection and classification in complex contexts [32]. Benchmarks for low-resolution text images highlight difficulties posed by blurred character shapes and detail loss, necessitating robust techniques for improved recognition.

The scarcity of pixel-level annotated datasets for scene text segmentation remains a significant hurdle, limiting effective deep learning model training [33]. Traditional OCR systems, relying heavily on CNNs and RNNs, often struggle with adaptability and performance across diverse scenarios [24].

Innovative approaches enhance STR systems' robustness and accuracy. Advanced computer vision techniques, like iterative rectification processes, address issues related to perspective distortion and text line curvature, improving recognition accuracy in natural scenes. The creation of specialized benchmarks and frameworks for low-resolution and low-resource languages emphasizes the need for adaptable and comprehensive STR solutions. This necessity is underscored by diverse text appearances and varying conditions in real-world scenarios, which traditional OCR methods struggle to manage effectively. Developing tailored tools allows researchers to better evaluate and enhance STR models, improving their performance across a broader range of applications, including complex text layouts and varied environmental factors [34, 35, 31, 36].

2.4 Multimodal Integration and Vision-Language Models

Multimodal integration is crucial for advancing vision-language models, enabling seamless processing and understanding of visual and textual information. Innovative techniques enhance semantic representation and comprehension across modalities. The Generative Cross-modal Learning Network (GXN) exemplifies this by embedding generative processes into cross-modal feature learning, capturing both global abstract and local grounded features to enhance representation learning [37].

Models like TextHarmony illustrate the potential of harmonizing vision and language generation within a single model, addressing challenges in integrating these modalities [6]. This approach underscores the importance of unified models that can process and generate coherent outputs across different data types.

Advanced frameworks categorizing data augmentation strategies using large language models (LLMs) distinguish between generative and discriminative learning approaches [2]. Such frameworks refine multimodal systems by systematically enhancing training data diversity and quality, crucial for robust vision-language model development.

The integration of sophisticated NLP tools leveraging deep learning is vital for advancing interactive language processing, as highlighted in studies focusing on language applications [3]. These tools facilitate nuanced language understanding alongside visual data, enabling accurate and contextually relevant outputs in tasks like image captioning and visual question answering.

Moreover, weakly supervised learning techniques that convert bounding-box annotations into pixel-level supervisions exemplify innovative approaches to improving segmentation models on real images [33]. This enhances data representation granularity, thereby improving multimodal models' performance in processing complex visual scenes.

The fusion of multimodal signals and the development of sophisticated representation learning techniques are fundamental to evolving vision-language models. By integrating visual and textual data through innovative methodologies, researchers can create robust and versatile models excelling in various applications, including image captioning, text-to-image generation, and visual question answering, ultimately advancing artificial intelligence [25].

3 Scene Text Synthesis

Category	Feature	Method
Methods and Techniques in Scene Text Synthesis	Resolution and Refinement	StackGAN[38], TG[39]
	Learning and Supervision Techniques	ET[40]
Advancements in Scene Text Synthesis	Text and Layout Generation	VRCIG[41], AT[42]
	Attention and Alignment Techniques	GAT[43], DT[44], XMC-GAN[45], ControlGAN[46]
	Model and Training Strategies	TGAN[47], TrOCR[24], SD-GAN[27], SMANet[33]
	Detail and Realism Enhancement	EnsNet[1], AttnGAN[48], ST[36]
Role of Generative Adversarial Networks	Semantic Coherence and Consistency	VT-Baseline[10], TediGAN[4], MG[49]
	Spatial and Layout Optimization	STT[50], GTR[51], HTIS[52], SF-GAN[53]
	Problem Decomposition and Integration	ST[54], TH[6]

Table 1: Table summarizing the key methods, advancements, and the role of Generative Adversarial Networks (GANs) in scene text synthesis. It categorizes various techniques and innovations, highlighting their contributions to enhancing visual realism, semantic coherence, and integration within images. The table provides a comprehensive overview of the methodologies that have shaped the current landscape of scene text synthesis.

AI-generated, for reference only.

In scene text synthesis, methodologies are crucial for generating visually coherent and contextually relevant text within images. This section examines innovative techniques that enhance the quality of synthesized text, analyzing their effectiveness and applications. Table 1 provides a detailed overview of the methods, advancements, and the role of Generative Adversarial Networks in the field of scene text synthesis, highlighting the diverse techniques and innovations that contribute to the generation of visually realistic and contextually coherent text within images. Additionally, Table 3 provides a detailed comparison of various methods employed in scene text synthesis, elucidating their unique features and primary focus areas. Figure 2 illustrates the hierarchical categorization of these methodologies, advancements, and the role of Generative Adversarial Networks (GANs) in scene text synthesis. The figure highlights key frameworks, techniques, and innovations that contribute to generating visually realistic and contextually coherent text within images, thereby providing a comprehensive overview of the current landscape in this field.

3.1 Methods and Techniques in Scene Text Synthesis

Scene text synthesis involves generating text within images that is both visually realistic and contextually coherent. The SwapText framework exemplifies this by using text swapping and background completion to maintain the original text style in images [54]. SynthTIGER improves scene text recognition by integrating target and noise text images, reflecting real-world text appearances [36].

Hierarchical Text-to-Image Synthesis (HTIS) adopts a two-stage approach, inferring a semantic layout from text and synthesizing images from that layout, enhancing structured text integration [52]. StackGAN generates low-resolution images first, refining them to high-resolution, addressing previous limitations [38].

Glyph-Aware Training (GAT) methods utilize mixed granularity input strategies to represent glyph words as whole units, improving text rendering fidelity through attention alignment loss and OCR recognition loss [43]. However, challenges remain in accurately capturing text glyph structures, leading to generation inaccuracies [39].

Multimodal frameworks are critical, as shown by methods converting structured text into scene graphs, predicting bounding boxes for entities, and adjusting them using relational units [41]. The AnyText framework introduces a text-control diffusion pipeline, producing high-quality visual text across multiple languages and styles [42].

Attention mechanisms in GAN-based models, such as AttnGAN, enhance detailed text-to-image generation by focusing on relevant words during synthesis [48]. EvoText's self-escalation learning process enables continuous learning from new data, employing discriminator models to validate generated text [40]. The SD-GAN framework distills semantic commons from text while preserving semantic diversity through enhanced visual-semantic embedding [27].

Techniques like MTRNet integrate text removal and synthesis processes, while SF-GAN synthesizes images by merging foreground objects with background images, maintaining realism in both geometry and appearance [8, 53]. SMANet employs weakly supervised learning to convert bounding-box annotations into pixel-level supervision, improving scene text segmentation [33].

These methodologies highlight the diverse techniques in scene text synthesis, emphasizing the integration of generative models, multimodal frameworks, and iterative refinement processes. Innovations such as the AnyText diffusion model and the unified TextHarmony model advance AI-generated imagery, enhancing the accuracy and coherence of embedded text while addressing challenges like blurred or illegible text. Large-scale datasets such as AnyWord-3M and DetailedTextCaps-100K facilitate rigorous evaluation and benchmarking, expanding the possibilities of creating contextually integrated and realistic text within images [6, 20, 42].

Figure 3 illustrates the hierarchical classification of methods and techniques in scene text synthesis, emphasizing three primary categories: generative models, multimodal frameworks, and iterative refinements. Each category includes specific methods that contribute to advancing the field of scene text synthesis. The first image showcases precision in character placement, while the second demonstrates the technique's versatility across various themes and styles, enhancing the visual richness of digital imagery [55, 44].

Method Name	Visual Realism	Semantic Coherence	Multimodal Integration
ST[54]	High Visual Realism	Semantic Coherence	Multimodal Integration
ST[36]	Realistic Text Images	Reflect Real-world Appearances	Combining Target Noise
HTIS[52]	Realistic Images	Semantically Meaningful Images	Text-to-image Synthesis
AT[42]	High-quality Visual	Contextually Coherent Text	Text-control Diffusion
GAT[43]	Aesthetically Pleasing Images	Coherent Visual Texts	Text-to-image Models
VRCIG[41]	Realistic And Coherent	Preserve Visual Relations	Text And Image
AttnGAN[48]	Fine-grained Details	Text Description Alignment	Word-context Vectors
XMC-GAN[45]	Photo-realistic Scenes	High Semantic Fidelity	Contrastive Learning
ControlGAN[46]	High-quality Images	Semantically Match Text	Text-to-image
TGAN[47]	Realistic Sentences	Semantic Coherence	Adversarial Training
STT[50]	Text-focused Images	Text Recognition Accuracy	Image-text Matching
EnsNet[1]	Realistic Background Generation	Semantically Aligned	End-to-end Architecture
DT[44]	Image Quality Measures	Semantically Aligned Descriptions	Multimodal Integration
SD-GAN[27]	Photo-realistic Images	Semantic Consistency	Multimodal Integration
SMANet[33]	Natural Images	Text Segmentation Accuracy	Text Segmentation Network
TrOCR[24]	Visual Features Extraction	Semantically Aligned Generation	End-to-end Integration

Table 2: This table presents a comparative analysis of various scene text synthesis methods, focusing on three key attributes: visual realism, semantic coherence, and multimodal integration. The methods are evaluated based on their ability to generate realistic text images, maintain semantic alignment with text descriptions, and effectively integrate text and visual elements. The table highlights the diverse approaches and innovations in the field, providing insights into the strengths and capabilities of each method.

3.2 Advancements in Scene Text Synthesis

Recent advancements in scene text synthesis have significantly improved the generation of text within images, emphasizing visual realism and semantic coherence. Techniques like SwapText ensure stylistic consistency through text swapping and background completion [54]. SynthTIGER enhances realism by incorporating noise from surrounding text regions [36].

HTIS aids in generating semantically meaningful images by decomposing synthesis into semantic layout inference and image generation stages, aligning with complex text descriptions [52]. The AnyText framework excels in multilingual text generation, demonstrating superior performance across diverse linguistic contexts [42].

Innovations such as Feature-Aware Conditional GAN (FA-GAN) combine adversarial training with feature and category awareness, enhancing output realism and diversity [43]. Graph convolutional networks for textual reasoning leverage spatial context, marking an advancement over methods reliant solely on linguistic context [41].

Studies using datasets like MSCOCO and Flickr30K have optimized natural language generation, achieving high-quality image and video outputs with increased diversity and photorealism. AttnGAN's ability to leverage both global sentence-level and fine-grained word-level information allows for detailed image generation that accurately reflects text descriptions [48]. XMC-GAN produces higher quality images aligning with detailed descriptions while simplifying model architecture [45].

ControlGAN introduces word-level spatial and channel-wise attention-driven generators, enhancing control and quality of synthesized text [46]. TextGAN advances GAN-based text generation, furthering realistic sentence generation [47].

Transformer-Based Super-Resolution Networks (TBSRN) in scene text synthesis, exemplified by STT, emphasize character details and positions, improving text clarity [50]. EnsNet's lateral connection structure and refined loss function enhance detail capture and background reconstruction [1]. DreamText outperforms state-of-the-art methods in high-fidelity scene text synthesis, demonstrating effectiveness in both qualitative and quantitative evaluations [44].

The Siamese mechanism in frameworks like SD-GAN ensures semantic consistency and introduces semantic-conditioned batch normalization, enhancing semantic diversity in generated images [27]. Experiments with COCO TS and MLT S datasets show significant improvements in segmentation network generalization on real images, highlighting the effectiveness of integrating synthetic data with real-world applications [33]. TrOCR achieves state-of-the-art results in various text recognition tasks, showcasing the efficacy of pre-trained Transformers in scene text synthesis [24].

AI-generated, for reference only.

Advancements in text-to-image generation, particularly through Attentional Generative Adversarial Networks (AttnGAN), represent a leap in creating high-quality images aligned with natural language descriptions. This progress is characterized by attention-driven, multi-stage refinement processes, enabling models to focus on specific words in text descriptions for fine-grained detail generation. Innovations in multimodal integration techniques, including deep attentional multimodal similarity models, enhance image-text matching accuracy, leading to substantial improvements in performance metrics on challenging datasets. These advancements reflect a broader trend in multimodal learning, where the interplay between vision and language continues to evolve, facilitating contextually integrated and visually coherent outputs [22, 48]. Table 2 provides a comprehensive comparison of recent advancements in scene text synthesis, detailing the visual realism, semantic coherence, and multimodal integration capabilities of various state-of-the-art methods.

3.3 Role of Generative Adversarial Networks

Generative Adversarial Networks (GANs) have revolutionized scene text synthesis by enabling the generation of images that are visually realistic and semantically coherent. The GAN architecture, comprising a generator and a discriminator, fosters a dynamic adversarial training process that enhances the fidelity and diversity of synthesized text images. Models like TediGAN leverage the hierarchical characteristics of StyleGAN's latent space for meaningful manipulation while maintaining high-quality image synthesis [4].

Innovative frameworks such as SwapText utilize a divide-and-conquer strategy with specialized networks to effectively manage geometric distortions, improving text synthesis accuracy in complex visual contexts [54]. The Spatial Fusion GAN (SF-GAN) focuses on both geometry and appearance, ensuring synthesized images are indistinguishable from real images through adversarial training [53].

The layout generator in HTIS exemplifies advancements in generating semantically meaningful images, enhancing both output quality and interpretability [52]. MirrorGAN's global-local collaborative attention model significantly improves cross-domain semantic consistency, refining the generative process compared to existing methods [49].

Recent advancements have incorporated self-attention mechanisms, as seen in the Scene Text Transformer (STT), allowing models to focus on text regions while disregarding irrelevant background information, thus improving recognition accuracy [50]. Graph-based approaches, exemplified by GTR, refine character sequence predictions by leveraging spatial relationships, enhancing scene text synthesis performance [51].

The Slide-LoRA innovation dynamically aggregates modality-specific and modality-agnostic LoRA experts, offering a more integrated approach to multimodal generation and improving synthesis quality [6]. The cascaded approach introduced in recent studies integrates various state-of-the-art modules and design enhancements, significantly boosting visual translation and text synthesis performance [10].

GANs are essential for advancing scene text synthesis, providing sophisticated frameworks that enable high-quality text generation integrated contextually within images. These frameworks utilize attention mechanisms and conditional generation techniques to ensure synthesized text aligns accurately with visual elements while maintaining semantic coherence, enhancing applications in image editing, video generation, and multimodal learning [22, 46, 48]. Ongoing developments in GANs not only enhance computer vision and natural language processing capabilities but also lay the groundwork for future innovations in multimodal data synthesis and generation.

Feature	SwapText	SynthTIGER	Hierarchical Text-to-Image Synthesis (HTIS)
Technique Type	Text Swapping	Image Integration	Two-Stage Approach
Primary Focus	Stylistic Consistency	Realism Enhancement	Semantic Layout
Unique Feature	Background Completion	Noise Incorporation	Structured Integration

Table 3: This table presents a comparative analysis of three prominent methods in scene text synthesis: SwapText, SynthTIGER, and Hierarchical Text-to-Image Synthesis (HTIS). It highlights the distinctive features and primary focus of each technique, emphasizing their respective contributions to stylistic consistency, realism enhancement, and semantic layout in text synthesis.

4 Visual Text Generation

The evolution of visual text generation is marked by significant advancements in creating outputs that are both semantically rich and visually coherent, effectively bridging the gap between textual descriptions and their visual counterparts. This progress has been driven by innovative approaches and frameworks leveraging advanced machine learning models, which we explore in the following subsections, highlighting their contributions to enhancing the quality and effectiveness of visual text generation.

4.1 Approaches and Frameworks

Advancements in visual text generation have been propelled by approaches that integrate text and image modalities. SceneVTG, for instance, employs Multimodal Large Language Models (MLLMs) to synthesize realistic scene text images, ensuring visual realism and semantic coherence [56]. Stacking-GANs, utilized in visual-relation layout methods, enhance image fidelity and resolution by generating high-resolution images from structured text [41]. State-of-the-art models like DALL-E3, AnyText, and DS-Fusion demonstrate capabilities in producing diverse, high-quality visual content [57].

CogView leverages a powerful Transformer architecture to learn complex distributions and relationships between text and images, crucial for generating relevant visual outputs [58]. Meanwhile, MTRNet excels in partial text removal and multilingual generalization, aided by auxiliary masks for training stability [8]. These frameworks highlight advancements driven by GAN architectures, Transformer models, and multimodal integration techniques. Recent developments include hybrid frameworks combining Variational Autoencoders (VAEs) and GANs for capturing static and dynamic video elements from text, while SceneVTG addresses real-world text image generation challenges using MLLMs for fidelity and reasonability [6, 22, 56].

Figure 4 illustrates various methodologies in visual text generation, each contributing uniquely to Explainable AI (ExAII) and generative models. The flowchart categorizes diverse methods essential for elucidating AI decision-making processes. The second diagram highlights document structure, crucial for understanding organization and retrieval. The third focuses on adversarial training dynamics, essential for refining model outputs through iterative learning [59, 60, 40].

4.2 Coherence and Contextual Alignment

Ensuring coherence and contextual alignment in visual text generation is vital for producing semantically meaningful and visually integrated content. SceneVTG employs a two-stage process with a Text Region and Content Generator (TRCG) and a Local Visual Text Renderer (LVTR) to generate high-quality text images [56]. The Text-Animator framework maintains alignment in dynamic environments, enhancing user experience [61].

Frameworks like TextHarmony and CogView address modality-specific inconsistencies by harmonizing text and image generation, enhancing multimodal outputs' reliability and quality. TextHarmony uses dynamic aggregation of modality-specific experts, while Prompt-Consistency Image Generation (PCIG) leverages knowledge graphs and diffusion models to align visual outputs with textual descriptions [6, 9, 58]. Advanced models interpret and generate text sensitive to visual cues, refining visual text generation's fidelity and applicability.

4.3 Fidelity and Reasonability

Fidelity and reasonability are critical metrics for evaluating visual text generation systems. SceneVTG utilizes Multimodal Large Language Models (MLLMs) and a local conditional diffusion model to ensure coherent text generation seamlessly integrated with visual backgrounds [56]. Fidelity involves accurately reflecting intended semantics and aesthetics, while reasonability ensures contextual appropriateness within the visual environment.

Advanced frameworks enhance text generation fidelity and reasonability through attention models and conditional diffusion processes, facilitating precise text attribute manipulation. GlyphControl enhances text-to-image models by incorporating glyph conditional information, enabling effective content customization. Approaches combining VAEs with holistic attribute discriminators impose semantic structures for controlled text generation with specific attributes [5, 28, 39]. These technologies ensure visually appealing, contextually relevant text, enhancing visual text generation systems' effectiveness.

4.4 Innovations in Text Representation

Innovations in text representation for visual text generation significantly advance the field by integrating textual and visual modalities. GlyphControl introduces glyph conditional information to improve text-to-image model performance, ensuring fidelity and coherence [28]. Prompt-Consistency Image Generation (PCIG) refines text representation by integrating large language models, knowledge graphs, and controllable diffusion models to ensure consistency and mitigate hallucinations [9].

Attention mechanisms within GAN-based models, like AttnGAN, facilitate fine-grained text-to-image generation, focusing on relevant words for detailed outputs [48]. Multimodal frameworks convert structured text into scene graphs, emphasizing spatial and relational reasoning for accurate text placement [41]. Innovations like AnyText and TextHarmony enhance realistic, contextually integrated text generation. AnyText addresses multilingual text generation challenges using a sophisticated pipeline, while TextHarmony harmonizes visual text and image generation, overcoming modality inconsistencies with Slide-LoRA [6, 42].

5 Applications and Use Cases

5.1 Augmented Reality and Assistive Technologies

The integration of scene text synthesis and visual text generation has notably advanced augmented reality (AR) and assistive technologies, particularly benefiting visually impaired users by interpreting and presenting textual information in accessible formats. Utilizing natural language processing (NLP) and deep learning, these technologies achieve high accuracy through synthetic datasets like MJSynth and SynthText, alongside real-world datasets such as IC13, IC15, IIIT, and SVT [31]. In AR, scene text synthesis overlays contextual information onto real-world views, enhancing user interaction and understanding. TextGen's multilingual text generation capabilities further extend AR's usability across languages [39].

Assistive technologies convert visual text into accessible formats, with NLP extracting insights to provide personalized assistance [19]. These tools deliver real-time support, such as reading aloud environmental text or providing scene descriptions. The convergence of scene text synthesis and visual text generation in AR and assistive technologies enhances information accessibility and contextual relevance, fostering inclusive applications through multimodal intelligence [20, 23].

5.2 Autonomous Driving and Geographic Information Systems

The integration of scene text synthesis and visual text generation in autonomous driving and geographic information systems (GIS) marks a significant technological leap. In autonomous driving, multilingual scene text detection and recognition are vital for interpreting road signs and traffic signals, crucial for safe navigation [62]. Realistic text image synthesis enhances scene text recognition model training, enabling robust performance in diverse linguistic contexts.

For GIS, these technologies improve data extraction accuracy and efficiency from visual content, facilitating automatic updates and validation of geographic data, essential for maintaining accurate maps and spatial databases [62]. Deep learning and NLP advancements offer sophisticated models for processing complex visual and textual data, enhancing accuracy and contextual awareness in real-world environments [23, 63].

5.3 Social Media and Digital Communication

Text generation's role in social media and digital communication has surged, driven by NLP and deep learning advancements. AI-driven content generation technologies, particularly text-to-image and text-to-video models, create personalized, contextually relevant content, essential for user engagement in digital environments. Techniques like hierarchical feature mapping and variational autoencoders enhance editability and semantic accuracy, addressing user preferences and contextual nuances

[5, 20, 29, 64, 22]. Scene text synthesis and visual text generation produce visually appealing, semantically rich content, boosting user interaction.

In social media, text generation automates captions, comments, and responses, enhancing communication and capturing user attention. Recent multimodal generative models, including Variational Autoencoders and Generative Adversarial Networks, improve text-to-video generation, enriching user engagement through synchronized visual-textual narratives [6, 22, 29]. Generative models like GANs and Transformers synthesize high-quality visual content, complementing textual information and enriching user experience.

Digital communication platforms benefit from automated content creation and moderation tools, utilizing advanced NLP models and large pre-trained transformer architectures for contextually relevant, linguistically precise text generation, facilitating seamless communication across diverse user groups [19, 11, 12, 13, 21]. Large language models (LLMs) enhance personalized, context-aware message generation, crucial for user engagement and satisfaction.

Advancements in text generation technologies transform social media and digital communication, enabling dynamic, interactive content creation. By leveraging AI, deep learning, and NLP for text-to-image and text-to-video generation, platforms enhance user experiences, facilitating engaging content creation like image memes and promoting meaningful interactions across digital channels [20, 29].

5.4 AI Art and Commercial Applications

Text synthesis in AI art and commercial sectors has transformed content creation, employing advanced generative models for diverse, personalized outputs. In AI art, techniques like RTT-GAN enable unique description generation from single images through adversarial training, ensuring high-quality artistic outputs [65]. This capability enhances creativity, offering new avenues for artists to create visually compelling, contextually rich art.

In commercial applications, text synthesis addresses real-world data limitations, generating synthetic data to enhance machine learning applications across domains [66]. This strategy creates high-quality datasets for effective model training, improving accuracy and performance in commercial applications from marketing to product development.

The unification of generative and discriminative visual foundation models (VFMs) highlights their transformative potential in computer vision, enabling seamless text and visual data integration for aesthetically pleasing, functionally relevant content [67]. Ongoing research seeks to address challenges, ensuring text synthesis in AI art and commercial sectors remains innovative. Leveraging these advancements allows businesses to create engaging, personalized customer experiences, driving growth and success in the digital economy.

6 Challenges and Future Directions

The domain of text synthesis and generation presents intricate challenges that demand innovative solutions to overcome current methodological constraints and enable future progress.

6.1 Challenges in Current Methods

Text synthesis and generation methodologies are hindered by computational constraints, data limitations, and methodological issues. Training instability in Generative Adversarial Networks (GANs) often leads to mode collapse, affecting output diversity and fidelity [14]. Convolutional Neural Networks (CNNs) introduce image-specific biases that complicate Optical Character Recognition (OCR) [24]. Domain shifts between synthetic data and real images impede model generalization, especially for those trained mainly on synthetic datasets [33]. Limited availability of large-scale, manually labeled data restricts the development of robust models capable of diverse task generalization [25]. Furthermore, preserving the visual integrity of text and background during translation poses significant challenges [10]. Generating images from complex descriptions with varying semantic interpretations remains problematic [27]. Addressing these issues requires innovative solutions to tackle computational complexities, data limitations, and methodological shortcomings as Natural Language Processing (NLP) continues to evolve [21, 12].

6.2 Data Quality and Diversity

Effective text synthesis and generation models rely heavily on data quality and diversity. High-quality datasets are crucial for generalization across tasks, reducing overfitting risks, and enhancing output robustness. Comprehensive datasets enable coherent and contextually relevant text generation in applications like conversation generation, machine translation, and summarization [21, 68]. However, current methodologies face challenges due to limited data sources. Reliance on web-scale image-text pairs, as seen in models like LaVIT, can introduce noise and complicate the modeling of complex image-text relationships [69]. This dependence risks propagating biases from original datasets, impacting the fairness and representativeness of synthetic data [66]. The scarcity of large-scale, well-annotated image-text data, especially in non-Latin languages, constrains model accuracy in multilingual contexts [42]. Training data diversity is vital for addressing extreme text appearances that may be underrepresented in datasets. While methods like SynthTIGER enhance scene text recognition robustness, they struggle with certain extreme cases due to insufficient representation [36]. Similarly, the quality and quantity of training data in methods like StackGAN directly affect the diversity of generated images, underscoring the need for more comprehensive datasets [38]. Adapting NLP tools to specific linguistic features, such as Turkish, and managing the complexity of languages like Arabic present significant hurdles in creating diverse datasets [3, 70]. Innovative approaches like EvoText offer scalable solutions for continuous learning in generative models without altering their architecture, making them suitable for low-resource NLP tasks [40]. However, difficulties in handling complex scenes or abstract descriptions persist, necessitating advancements in data collection and model training techniques [45].

6.3 Model Generalization and Robustness

Model generalization and robustness are essential for effective text synthesis and generation across diverse datasets. Current methods struggle with complex visual and linguistic data, such as curved or occluded text, impacting output accuracy and coherence [10]. Future research should optimize computational efficiency and enhance model adaptability to various text complexities [71]. Integrating deep learning techniques, particularly Recurrent Neural Networks (RNNs) and CNNs, shows promise for improving extraction accuracy and robustness in text synthesis models [16]. Leveraging these architectures can enhance generalization across applications, improving performance in real-world scenarios. Optimizing computational efficiency in word importance quantification is crucial for extending current methods to more complex datasets, addressing limitations in translating longer sentences, and maintaining visual coherence, which are vital for ensuring reliability and accuracy in diverse contexts [10].

6.4 Technological and Architectural Advancements

The future of scene text synthesis and visual text generation is poised for significant advancements through the development of hybrid models that integrate diverse architectures, enhancing robustness and interpretability across domains. These hybrid approaches are essential for addressing challenges such as handling out-of-vocabulary (OOV) words and improving performance in various applications. Future research should explore enhancing model adaptability to diverse motion scenarios and optimizing generation efficiency [61]. Optimizing the speed of these methods and their application in video text generation and fine-grained object generation is critical for future research, significantly enhancing real-time applicability in dynamic environments [22]. Enhancing weakly supervised learning frameworks like WeText by incorporating diverse datasets and exploring iterative learning approaches could further improve model performance [32]. Expanding language support and improving evaluation protocols are vital for addressing challenges related to recognizing distorted text. The integration of foundation models in data synthesis presents a promising direction, with future research focusing on developing robust evaluation frameworks and mitigating biases in synthetic data [2]. Enhancing semantic representation of captions and exploring additional NLP techniques could further improve retrieval system robustness and accuracy [72]. Refining model efficiency and exploring additional synthesis tasks are crucial for enhancing generalization capabilities across domains [73]. Implementing advanced NLP algorithms like BERT, XLNet, and GPT-2 can improve knowledge summarization accuracy and efficiency in various fields, including material science. Future research should also enhance model performance on underrepresented languages and improve robustness against diverse text orientations and complexities [36]. Further advancements in conditioning methods and dataset expansion are essential for improving model robustness [38]. The development of hybrid architectures, hardware acceleration techniques, and optimized hyperparameter tuning processes are crucial for advancing deep learning frontiers [21]. Scaling datasets and transitioning computations to GPU systems can enhance performance and facilitate meaningful comparisons with classical models. Future research could also focus on improving the model's ability to maintain global coherence in generated images and integrating additional modalities to enhance image quality [48]. Additionally, improving the model's capacity for complex narratives and integrating modalities for better image generation are key areas for exploration [45]. Future research should enhance STT capabilities for complex text scenarios and improve performance on unseen styles [50]. Enhancements in handling diverse image contexts and optimizing speed and generalization for object removal tasks should be prioritized [1]. Further optimizations for multimodal integration and generating images with dense text are promising directions [6]. Improving attribute disentanglement in latent space and enhancing robustness against out-of-distribution images could significantly advance the field [4]. Future directions include refining training techniques, enhancing performance on diverse datasets, and investigating MTRNet applicability beyond text removal [8]. These technological and architectural advancements have the potential to significantly enhance scene text synthesis and visual text generation capabilities, driving innovation and expanding applicability across various domains. Future research should focus on improving controllability of text attributes, expanding multi-language support, and developing end-to-end pipelines for real-world applications [57].

7 Conclusion

The integration of deep learning, generative adversarial networks, and optical character recognition has significantly propelled the field of scene text synthesis and visual text generation, enabling the creation of text within images that is both realistic and contextually relevant. This progress has bolstered applications across various domains, including augmented reality, autonomous driving, and digital communication, by enhancing the interaction between textual and visual data. The role of multi-modal representation learning is pivotal, as it facilitates the seamless integration of these data types, thereby improving the quality and applicability of generated content.

Despite these advancements, challenges persist, particularly in the areas of model generalization and robustness when applied to diverse datasets and environments. The scarcity of high-quality, varied datasets remains a critical obstacle to achieving effective model generalization. Innovative strategies, such as data augmentation through large language models, offer promising solutions for overcoming these challenges and fostering the development of new learning paradigms that could drive future progress.

In the context of natural language processing, especially for languages with limited resources like Arabic, there are ongoing challenges related to resource availability and the creation of comprehensive tools. Addressing these issues is crucial for advancing NLP applications and enhancing their performance across different linguistic contexts.

Looking ahead, the future of scene text synthesis and visual text generation is promising, with the potential for more advanced models capable of handling complex visual and linguistic data. Continued research and innovation are expected to lead to substantial improvements in accuracy, efficiency, and overall applicability, thereby expanding the influence of these technologies across a broader range of applications.

References

[1] Shuaitao Zhang, Yuliang Liu, Lianwen Jin, Yaoxiong Huang, and Songxuan Lai. Ensnet: Ensconce text in the wild. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 801–808, 2019.

[2] Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. Data augmentation using large language models: Data perspectives, learning paradigms and challenges, 2024.

[3] Kadir Tohma and Yakup Kutlu. Challenges encountered in turkish natural language processing studies, 2021.

[4] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021.

[5] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR, 2017.

[6] Zhen Zhao, Jingqun Tang, Binghong Wu, Chunhui Lin, Shu Wei, Hao Liu, Xin Tan, Zhizhong Zhang, Can Huang, and Yuan Xie. Harmonizing visual text comprehension and generation, 2024.

[7] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

[8] Osman Tursun, Rui Zeng, Simon Denman, Sabesan Sivapalan, Sridha Sridharan, and Clinton Fookes. Mtrnet: A generic scene text eraser. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 39–44. IEEE, 2019.

[9] Yichen Sun, Zhixuan Chu, Zhan Qin, and Kui Ren. Prompt-consistency image generation (pcig): A unified framework integrating llms, knowledge graphs, and controllable diffusion models, 2024.

[10] Shreyas Vaidya, Arvind Kumar Sharma, Prajwal Gatti, and Anand Mishra. Show me the world in my language: Establishing the first baseline for scene-text to scene-text translation, 2024.

[11] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[12] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744, 2023.

[13] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

[14] He Huang, Philip S. Yu, and Changhu Wang. An introduction to image synthesis with generative adversarial nets, 2018.

[15] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[16] Firhan Maulana Rusli, Kevin Akbar Adhiguna, and Hendy Irawan. Indonesian id card extractor using optical character recognition and natural language post-processing, 2020.

[17] Yuyu Luo, Jiawei Tang, and Guoliang Li. nvbench: A large-scale synthesized dataset for cross-domain natural language to visualization task, 2021.

[18] Haonan Yu, Sergey Edunov, Yuandong Tian, and Ari S. Morcos. Playing the lottery with rewards and multiple languages: lottery tickets in rl and nlp, 2020.

[19] Yue Kang, Zhao Cai, Chee-Wee Tan, Qian Huang, and Hefu Liu. Natural language processing (nlp) in management research: A literature review. *Journal of Management Analytics*, 7(2):139–172, 2020.

[20] Aditi Singh. A survey of ai text-to-image and ai text-to-video generators, 2023.

[21] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9):1–39, 2024.

[22] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[23] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications, 2020.

[24] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13094–13102, 2023.

[25] Feng Li, Hao Zhang, Yi-Fan Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, PengChuan Zhang, and Lei Zhang. Vision-language intelligence: Tasks, representation learning, and large models, 2022.

[26] Sandeep Subramanian, Sai Rajeswar Mudumba, Alessandro Sordoni, Adam Trischler, Aaron C Courville, and Chris Pal. Towards text generation with adversarially learned neural outlines. *Advances in Neural Information Processing Systems*, 31, 2018.

[27] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2327–2336, 2019.

[28] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation, 2023.

[29] Zhiyuan Liu, Chuanzheng Sun, Yuxin Jiang, Shiqi Jiang, and Mei Ming. Multi-modal application: Image memes generation, 2021.

[30] Fei Yin, Yi-Chao Wu, Xu-Yao Zhang, and Cheng-Lin Liu. Scene text recognition with sliding convolutional character models. *arXiv preprint arXiv:1709.01727*, 2017.

[31] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4715–4723, 2019.

[32] Shangxuan Tian, Shijian Lu, and Chongshou Li. Wetext: Scene text detection under weak supervision. In *Proceedings of the IEEE international conference on computer vision*, pages 1492–1500, 2017.

[33] Simone Bonechi, Monica Bianchini, Franco Scarselli, and Paolo Andreini. Weak supervision for generating pixel–level annotations in scene text segmentation. *Pattern Recognition Letters*, 138:1–7, 2020.

[34] Pablo Pino, Denis Parra, Pablo Messina, Cecilia Besa, and Sergio Uribe. Inspecting state of the art performance and nlp metrics in image-based medical report generation, 2022.

[35] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European conference on computer vision*, pages 178–196. Springer, 2022.

[36] Moonbin Yim, Yoonsik Kim, Han-Cheol Cho, and Sungrae Park. Synthtiger: Synthetic text image generator towards better text recognition models. In *International conference on document analysis and recognition*, pages 109–124. Springer, 2021.

[37] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7181–7189, 2018.

[38] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.

[39] Boqiang Zhang, Zuan Gao, Yadong Qu, and Hongtao Xie. How control information influences multilingual text image generation and editing?, 2024.

[40] Zhengqing Yuan, Huiwen Xue, Chao Zhang, and Yongming Liu. Evotext: Enhancing natural language generation models via self-escalation learning for up-to-date knowledge and improved performance, 2023.

[41] Duc Minh Vo and Akihiro Sugimoto. Visual-relation conscious image generation from structured-text. In *European conference on computer vision*, pages 290–306. Springer, 2020.

[42] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing, 2024.

[43] Wenbo Li, Guohao Li, Zhibin Lan, Xue Xu, Wanru Zhuang, Jiachen Liu, Xinyan Xiao, and Jinsong Su. Empowering backbone models for visual text generation with input granularity control and glyph-aware training, 2024.

[44] Yibin Wang, Weizhong Zhang, and Cheng Jin. Dreamtext: High fidelity scene text synthesis, 2024.

[45] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021.

[46] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in neural information processing systems*, 32, 2019.

[47] Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. Adversarial feature matching for text generation. In *International conference on machine learning*, pages 4006–4015. PMLR, 2017.

[48] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[49] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1505–1514, 2019.

[50] Jingye Chen, Bin Li, and Xiangyang Xue. Scene text telescope: Text-focused scene image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2021.

[51] Yue He, Chen Chen, Jing Zhang, Juhua Liu, Fengxiang He, Chaoyue Wang, and Bo Du. Visual semantics allow for textual reasoning better in scene text recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 888–896, 2022.

[52] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7986–7994, 2018.

AI-generated, for reference only.

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

[54] Qiangpeng Yang, Jun Huang, and Wei Lin. Swaptext: Image based texts transfer in scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14700–14709, 2020.

[55] Fan Bai, Zhanzhan Cheng, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Edit probability for scene text recognition. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1508–1516, 2018.

[56] Yuanzhi Zhu, Jiawei Liu, Feiyu Gao, Wenyu Liu, Xinggang Wang, Peng Wang, Fei Huang, Cong Yao, and Zhibo Yang. Visual text generation in the wild, 2024.

[57] Yejin Choi, Jiwan Chung, Sumin Shim, Giyeong Oh, and Youngjae Yu. Towards visual text design transfer across languages, 2024.

[58] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021.

[59] Julia El Zini and Mariette Awad. On the explainability of natural language processing deep models, 2022.

[60] David Elworthy. Retrieval from captioned image databases using natural language processing, 2000.

[61] Lin Liu, Quande Liu, Shengju Qian, Yuan Zhou, Wengang Zhou, Houqiang Li, Lingxi Xie, and Qi Tian. Text-animator: Controllable visual text video generation, 2024.

[62] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khlif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, et al. Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1582–1587. IEEE, 2019.

[63] Rui Xie. Frontiers of deep learning: From novel application to real-world deployment, 2024.

[64] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023.

[65] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE international conference on computer vision*, pages 3362–3371, 2017.

[66] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. Machine learning for synthetic data generation: A review, 2024.

[67] Xu Liu, Tong Zhou, Yuanxin Wang, Yuping Wang, Qinjingwen Cao, Weizhi Du, Yonghuan Yang, Junjun He, Yu Qiao, and Yiqing Shen. Towards the unification of generative and discriminative visual foundation model: A survey, 2023.

[68] Sanskar Soni, Satyendra Singh Chouhan, and Santosh Singh Rathore. Textconvonet:a convolutional neural network based architecture for text classification, 2022.

[69] Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, Chengru Song, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023.

[70] Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Samhaa R. El-Beltagy, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. A panoramic survey of natural language processing in the arab world, 2021.

[71] Claudio Fanconi, Moritz Vandenhirtz, Severin Husmann, and Julia E. Vogt. This reads like that: Deep learning for interpretable natural language processing, 2023.

[72] Sarvesh Patil. Deep learning based natural language processing for end to end speech translation, 2018.

[73] Dewayne Whitfield. Using gpt-2 to create synthetic data to improve the prediction performance of nlp machine learning classification models, 2021.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.

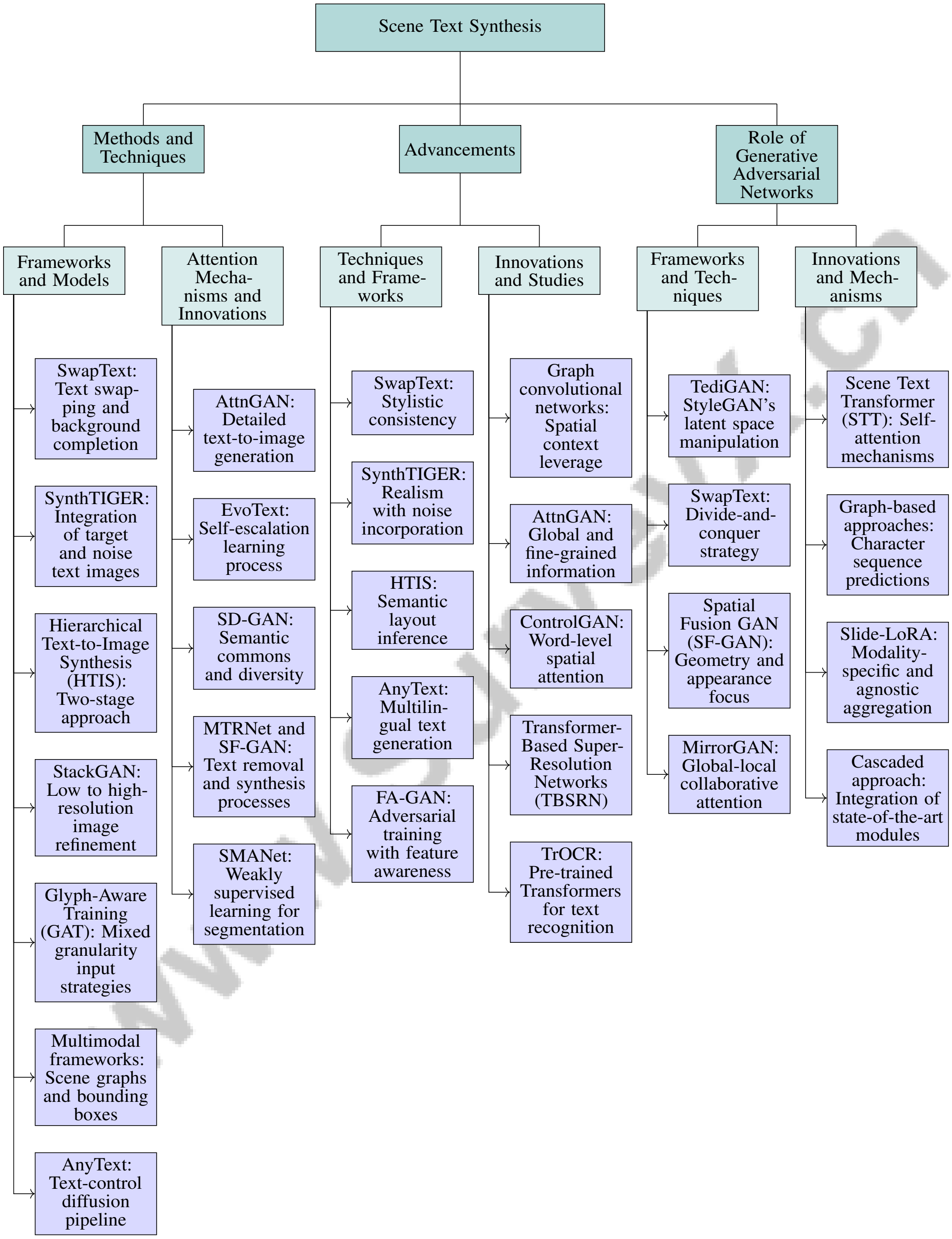

Scene Text Synthesis
Methods and Techniques
Advancements
Role of Generative Adversarial Networks
Frameworks and Models
Attention Mechanisms and Innovations
Techniques and Frameworks
Innovations and Studies
Frameworks and Techniques
Innovations and Mechanisms
SwapText: Text swapping and background completion
SynthTIGER: Integration of target and noise text images
Hierarchical Text-to-Image Synthesis (HTIS): Two-stage approach
StackGAN: Low to high-resolution image refinement
Glyph-Aware Training (GAT): Mixed granularity input strategies
Multimodal frameworks: Scene graphs and bounding boxes
AnyText: Text-control diffusion pipeline
AttnGAN: Detailed text-to-image generation
EvoText: Self-escalation learning process
SD-GAN: Semantic commons and diversity
MTRNet and SF-GAN: Text removal and synthesis processes
SMANet: Weakly supervised learning for segmentation
SwapText: Stylistic consistency
SynthTIGER: Realism with noise incorporation
HTIS: Semantic layout inference
AnyText: Multilingual text generation
FA-GAN: Adversarial training with feature awareness
Graph convolutional networks: Spatial context leverage
AttnGAN: Global and fine-grained information
ControlGAN: Word-level spatial attention
Transformer-Based Super-Resolution Networks (TBSRN)
TrOCR: Pre-trained Transformers for text recognition
TediGAN: StyleGAN's latent space manipulation
SwapText: Divide-and-conquer strategy
Spatial Fusion GAN (SF-GAN): Geometry and appearance focus
MirrorGAN: Global-local collaborative attention
Scene Text Transformer (STT): Self-attention mechanisms
Graph-based approaches: Character sequence predictions
Slide-LoRA: Modality-specific and agnostic aggregation
Cascaded approach: Integration of state-of-the-art modules

Figure 2: This figure illustrates the hierarchical categorization of methodologies, advancements, and the role of GANs in scene text synthesis, highlighting key frameworks, techniques, and innovations that contribute to generating visually realistic and contextually coherent text within images.

AI-generated, for reference only.

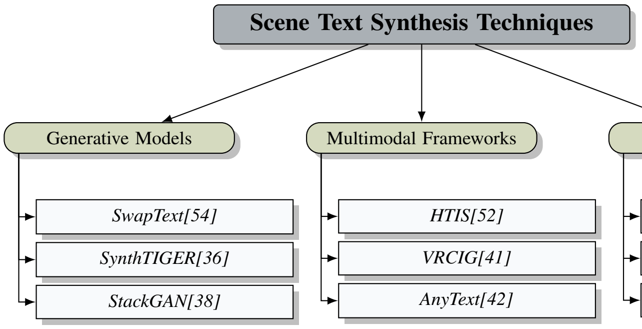
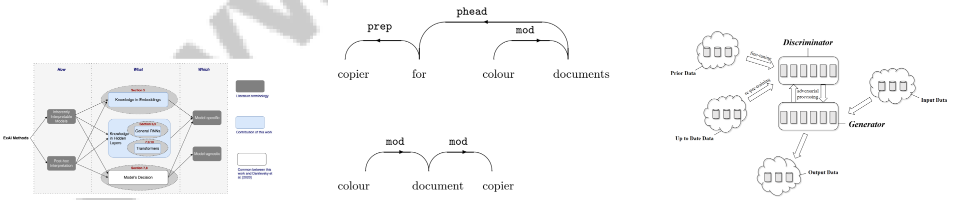


Figure 3: This figure illustrates the hierarchical classification of methods and techniques in scene text synthesis, emphasizing three primary categories: generative models, multimodal frameworks, and iterative refinements. Each category includes specific methods that contribute to advancing the field of scene text synthesis.



(a) The image depicts a flowchart illustrating the relationship between different methods and their contributions to the field of ExAII (Explainable AI).[59]

(b) A diagram illustrating the relationships between different parts of a document[60]

(c) Adversarial Training for Generative Models[40]

Figure 4: Examples of Approaches and Frameworks

AI-generated, for reference only.