A Survey of Diffusion Models in Computer Vision: Object Detection, Recognition, and Image Segmentation

Abstract

Diffusion models have emerged as a transformative technology in computer vision, significantly advancing the fields of image synthesis, object detection, recognition, and image segmentation. These models excel in generating high-fidelity data representations through iterative refinement processes, surpassing traditional generative models like GANs in various applications. This survey provides a comprehensive overview of diffusion models, examining their integration with other generative frameworks to enhance synthesis quality and efficiency. Key advancements include the development of neuroexplicit diffusion models and the integration of diffusion models with transformer architectures, which have improved performance across diverse vision tasks. In medical imaging, diffusion models have set new benchmarks, particularly in segmentation accuracy and efficiency. Despite their strengths, challenges such as computational demands and data quality persist. Proposed solutions focus on optimizing training algorithms and exploring novel model architectures. The survey highlights the synergies among diffusion models, object detection, recognition, and segmentation, emphasizing their role in driving innovation in computer vision. Future research directions include enhancing computational efficiency, exploring broader applications in medical imaging, and improving adversarial robustness. As diffusion models continue to evolve, they promise to unlock new possibilities in image synthesis, editing, and beyond, solidifying their position as a cornerstone of modern computer vision methodologies.

1 Introduction

1.1 Overview of Diffusion Models

Diffusion models have emerged as a critical element in generative modeling, particularly in computer vision, where they effectively transform random noise into coherent, high-quality data representations through iterative refinement. Utilizing a Markovian framework, these models progressively convert data into Gaussian noise and back, enabling the generation of images with remarkable fidelity [1]. The latent space of diffusion models, although more complex than that of Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), enhances their ability to capture the intricate structures of natural data.

The theoretical foundations of diffusion models are continually evolving, addressing knowledge gaps and expanding their applications [2]. Recent advancements highlight their effectiveness in high-resolution image synthesis, surpassing traditional super-resolution techniques and other generative models like GANs. For instance, diffusion models have significantly improved microscopy image resolution, demonstrating their utility in specialized fields by overcoming existing inefficiencies [3].

Beyond image synthesis, diffusion models exhibit versatility in video generation and editing, integrating generative and recognition capabilities. This adaptability extends to innovative applications,

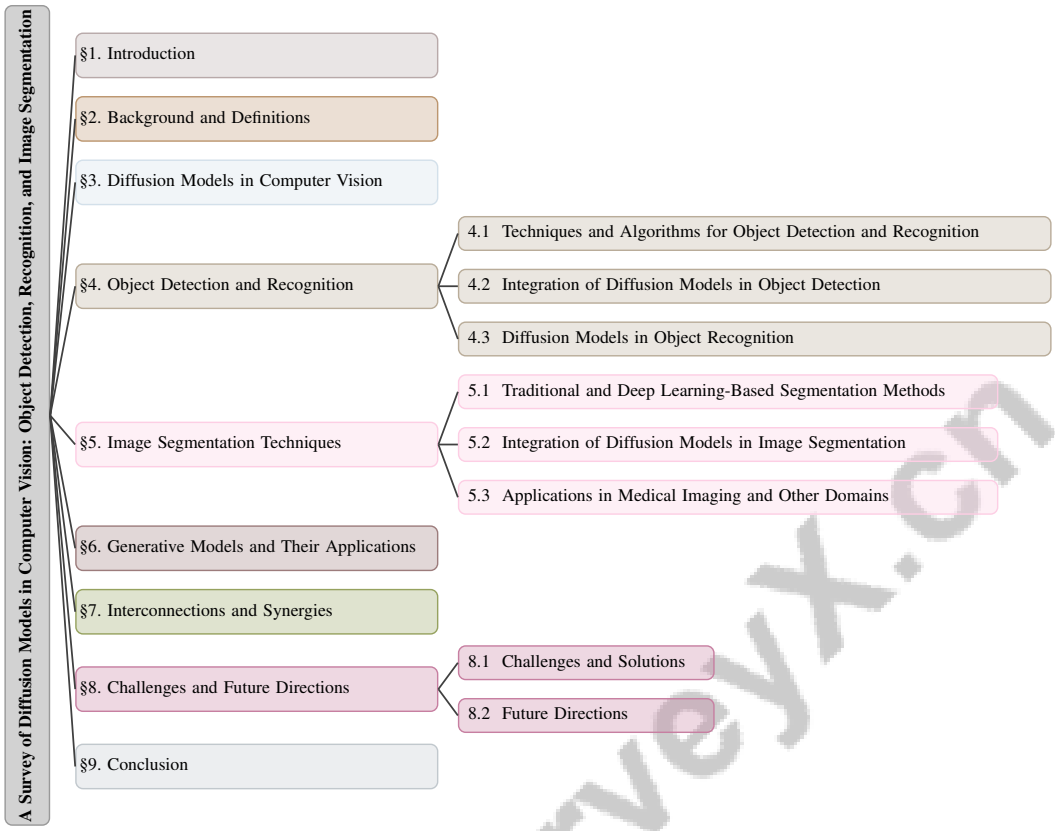


Figure 1: chapter structure

such as zero-shot translation in neural machine translation, showcasing their wide-ranging applicability [4]. In healthcare, diffusion models have shown promise in medical imaging tasks, including image-to-image translation, marking substantial advancements in real-world applications [5].

The rapid proliferation of diffusion models has spurred the development of new generative tools with potential commercial applications, raising concerns about output originality and ethical implications related to data memorization [6]. As research progresses, the role of diffusion models in advancing generative AI becomes increasingly evident, particularly in addressing challenges such as domain adaptation and the integration of vision-language tasks [7]. These models are poised to enhance various computer vision applications, including autonomous driving systems, by generating diverse and realistic scenarios that improve simulation coverage and realism [8]. The ongoing evolution of diffusion models indicates a promising trajectory for future exploration in computer vision.

1.2 Structure of the Survey

This survey is systematically structured to provide a comprehensive understanding of diffusion models within computer vision, focusing on their applications in object detection, recognition, and image segmentation. The initial sections introduce the interconnected areas of diffusion models, object detection, recognition, computer vision, image segmentation, and generative models, underscoring their significance in enabling computers to interpret and manipulate visual information.

Following this introduction, the survey presents foundational concepts and definitions essential for grasping the complexities of generative modeling, particularly in food imagery, where unique challenges such as intricate compositions and regional biases are examined. This foundational knowledge is critical for exploring sophisticated applications, including the manipulation of semantic concepts in diffusion models and the development of innovative training data generation techniques for object detection and segmentation [9, 10, 11]. The role of diffusion models in computer vision is then explored, emphasizing their capacity to generate data through iterative refinement processes and their transformative impact on the field.

AI-generated, for reference only.

Subsequent sections delve into specific applications, beginning with advancements in diffusion models and their roles in image generation and editing. These sections illustrate how diffusion models enhance diversity and fidelity in synthesis and editing tasks.

The survey then transitions to object detection and recognition, discussing both traditional and modern techniques, as well as the contributions of diffusion models to these areas. A comprehensive analysis of various image segmentation techniques follows, particularly focusing on the integration of diffusion models. This analysis evaluates how diffusion models improve segmentation accuracy and efficiency, leveraging their advanced capabilities to generate realistic and diverse image representations while aligning outputs with contextual instructions. Empirical studies supporting their competitive performance against traditional methods and specialized models are also highlighted [12, 13, 14, 15, 16].

The discussion emphasizes significant advancements brought about by generative models, particularly diffusion models, which enhance the quality and efficiency of synthesis processes. Operating by transforming data distributions into noise and reconstructing them, diffusion models achieve superior image synthesis capabilities across various applications. Notably, innovations such as retrieval-augmented diffusion models utilize external databases to improve generative performance, enabling effective scene composition without extensive paired training data. Additionally, optimizing inference diffusion processes through auxiliary variables enhances sample quality, facilitating experimentation and adaptation to diverse datasets. These developments represent a substantial leap in generative AI, offering researchers new tools and methodologies for advancing image generation and related tasks [17, 18, 2, 19]. The survey also investigates innovative frameworks and techniques that leverage synergies among these technologies, emphasizing their interconnections.

Finally, the survey examines the integration and application of diffusion models in computer vision, addressing significant challenges such as scalability and computational demands. It proposes future research directions aimed at improving model efficiency, enhancing controllability, and expanding applicability across diverse tasks, including video synthesis, image editing, and data augmentation [20, 14, 19]. The conclusion synthesizes the key points discussed, highlighting the importance of diffusion models in advancing computer vision technologies.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Diffusion Models: Foundations and Applications

Diffusion models have transformed generative modeling by iteratively refining noise into high-quality data representations. Denoising Diffusion Probabilistic Models (DDPMs) exemplify this, employing a Markovian framework to generate high-fidelity images through phase transitions [21]. Latent Diffusion Models (LDMs) further advance the field by enabling efficient training and high-resolution image sampling within a learned latent space [22].

Fourier Diffusion Models (FDMs) illustrate the versatility of diffusion models by utilizing linear shift-invariant systems and additive stationary Gaussian noise for stochastic image generation [23]. However, challenges such as encoding latent codes and computational demands during sampling persist [24, 25]. Innovations like integrating Transformer backbones instead of traditional UNet architectures enhance multimodal data processing efficiency [26]. Neuroexplicit diffusion models, combining explicit PDE-based diffusion with neural network parameterization, offer superior reconstruction and generalization [27].

In computer vision, diffusion models automate tasks like semantic segmentation. Techniques such as DiffuMask leverage text-guided diffusion models for high-resolution image generation with pixel-level semantic annotations [28]. These models incorporate various conditioning modalities, like segmentation maps and sketches, into pretrained models without extensive retraining [29].

Diffusion models also excel in modeling multi-modal distributions. The Gaussian Mixture Categorical Diffusion (GMCD) model uses diffusion in continuous space to encode categorical data, generating structurally coherent outputs [30]. ShiftDDPMs enhance flexibility by incorporating conditions into the forward process across all timesteps [31].

The foundational significance of diffusion models in computer vision is evident in specialized domains like microscopy image enhancement [3] and high-quality dMRI data generation from low-quality sources [32]. The ongoing evolution of diffusion models addresses limitations and explores new applications, solidifying their role in modern computer vision methodologies. Their utility and transformative potential are exemplified in tasks such as high-accuracy dichotomous image segmentation [33] and motion planning in high-dimensional scenarios [34].

2.2 Computer Vision and Image Segmentation

Computer vision, an interdisciplinary field merging computer science, neuroscience, and engineering, enables machines to analyze and interpret visual data. It encompasses computational models like deep learning architectures for feature extraction, probabilistic frameworks for scene interpretation, and segmentation techniques applicable in sectors such as healthcare, robotics, and transportation. Recent research highlights the impact of background signals on object recognition accuracy, enhancing machine capabilities in complex visual environments [35, 36, 37, 38, 10]. Key tasks include object detection, recognition, and image segmentation, essential for effective visual information processing.

The integration of diffusion models into computer vision, particularly for image segmentation, shows promise. These models, as powerful representation learners, address challenges like limited labeled data in semantic segmentation tasks [39]. By leveraging iterative refinement processes, diffusion models generate high-fidelity data representations critical for accurate segmentation outcomes.

In medical imaging, segmentation is crucial for applications like diagnosis and image-guided surgery. Traditional methods face challenges related to ambiguity and noise. Diffusion-based approaches, as discussed in [40], offer robust solutions by enhancing segmentation accuracy through improved training strategies [41]. These models exploit denoising capabilities to yield clearer, more precise segmentations, addressing conventional techniques' limitations.

Diffusion models also extend to unconventional domains like complex food imagery generation. The study of Concept Algebra in food image generation reveals culinary diversity and regional biases, highlighting diffusion models' potential to navigate these challenges and enhance image manipulation capabilities [11].

The incorporation of diffusion models into computer vision and image segmentation enhances accuracy and efficiency, broadening applicability across domains like image data augmentation, cross-domain semantic segmentation, and text-to-image synthesis. These models excel in generating realistic images by learning complex visual-semantic relationships and capturing universal features, bolstering machine learning models' robustness. Recent advancements focus on improving computational efficiency, facilitating real-world applications and addressing challenges related to energy consumption and model complexity [42, 43, 18, 14, 15]. By harnessing diffusion models' strengths, researchers achieve more reliable and nuanced interpretations of visual data, paving the way for advancements in theoretical and applied aspects of computer vision.

3 Diffusion Models in Computer Vision

Diffusion models have become a cornerstone in computer vision, transforming image synthesis and manipulation with their capacity to produce high-quality, diverse outputs. Figure 2 illustrates the hierarchical structure of diffusion models, highlighting key advancements such as neuroexplicit diffusion models, improvements in training efficiency, and their integration with vision transformers. This figure also covers various applications in image generation and editing, showcasing innovative techniques and theoretical frameworks that enhance visual quality and expand applicability. This section examines the latest advancements in diffusion models, emphasizing their enhanced performance and expanded applicability across various computer vision tasks. The following subsection will delve into these advancements, elucidating their implications for the future of generative modeling.

3.1 Advancements in Diffusion Models

Recent innovations in diffusion models have significantly improved their performance and applicability in computer vision. Notably, neuroexplicit diffusion models, which integrate explicit PDE-based

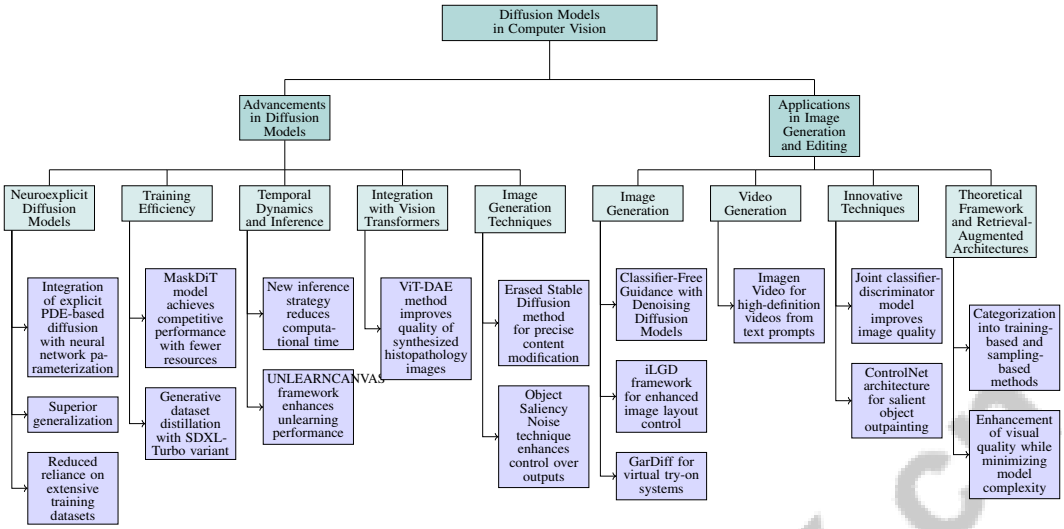


Figure 2: This figure illustrates the hierarchical structure of diffusion models in computer vision, highlighting advancements such as neuroexplicit diffusion models, training efficiency improvements, and integration with vision transformers. It also covers applications in image generation and editing, including innovative techniques and theoretical frameworks that enhance visual quality and expand applicability.

diffusion with neural network parameterization, offer superior generalization and reduced reliance on extensive training datasets, achieving high-quality outputs with fewer parameters [27].

Training efficiency has been enhanced by the MaskDiT model, which achieves competitive performance with fewer resources [44]. A novel generative dataset distillation method using the SDXL-Turbo variant of Stable Diffusion further improves image quality and increases images per class (IPC) [45].

Exploring temporal dynamics within diffusion models has led to a new inference strategy that reduces computational time while maintaining output quality by identifying significant generative phase patterns [46]. The UNLEARNCANVAS framework introduces dual supervision of styles and objects, enhancing unlearning performance [47].

Integration with vision transformers, as in the ViT-DAE method, has improved the quality of synthesized histopathology images [48], showcasing the versatility of diffusion models in handling complex data types.

In image generation, the Erased Stable Diffusion (ESD) method fine-tunes model parameters to erase specific visual concepts, demonstrating precision in modifying generated content [49]. The Object Saliency Noise (OSN) technique enhances control over outputs by embedding salient features of target images into input noise, improving fidelity and relevance [50].

These advancements underscore the transformative impact of diffusion models in computer vision, reinforcing their foundational role in the evolution of generative modeling technologies. As diffusion models progress, they promise to revolutionize image synthesis and editing by integrating innovative architectures and retrieval strategies that enhance visual quality and expand their applicability across various tasks, including text-guided generation and data augmentation [14, 18, 19].

3.2 Applications in Image Generation and Editing

Diffusion models have significantly advanced image generation and editing, improving both the diversity and fidelity of generated content. The integration of Classifier-Free Guidance with Denoising Diffusion Models has enabled the creation of highly realistic images, such as cataract surgery depictions, which are nearly indistinguishable from actual surgical images, highlighting their potential in generating high-quality medical imagery for training and educational purposes [51].

The iLGD framework enhances image layout control by combining attention injection with loss guidance, achieving superior image quality and layout precision [52]. In virtual try-on systems, GarDiff improves alignment with garment details by integrating a garment-focused adapter into the diffusion model [53].

In video generation, Imagen Video utilizes a cascade of video diffusion models to produce high-definition videos from text prompts, demonstrating the adaptability of diffusion models in handling dynamic visual content [54].

Innovative techniques, such as the joint classifier-discriminator model, improve image quality by combining classification with energy-based discrimination [55]. The OSN technique enhances control over outputs by embedding salient features into input noise, improving localization and saliency [50].

The ControlNet architecture adapts diffusion models for salient object outpainting, focusing on background generation while preserving object identity [56]. The ViT-DAE method leverages vision transformers to capture complex spatial layouts, significantly improving synthesized image quality [48].

The continuous evolution of diffusion models in image generation and editing emphasizes their transformative impact and versatility across applications, including text-to-image synthesis, video generation, and localized image manipulation. This advancement is supported by a robust theoretical framework that categorizes diffusion approaches into training-based and sampling-based methods, facilitating further research and innovation. Recent developments have introduced retrieval-augmented architectures that enhance visual quality while minimizing model complexity, enabling efficient generative capabilities even in novel domains [57, 18, 2, 19]. As diffusion models advance, they promise to expand the horizons of visual creativity and computational efficiency, driving further innovation in computer vision.

As depicted in Figure 3, this figure illustrates the hierarchical categorization of applications in image generation and editing, highlighting key methods and innovations such as Classifier-Free Guidance, the iLGD Framework, and GarDiff for virtual try-on in image generation; Imagen Video and the Joint Classifier-Discriminator in video generation; and Object Saliency Noise, ControlNet for outpainting, and the ViT-DAE Method in image manipulation. These visual aids underscore the progression and technical intricacies within the domain, emphasizing the transformative impact of diffusion models in enhancing image generation and editing processes [19, 58].

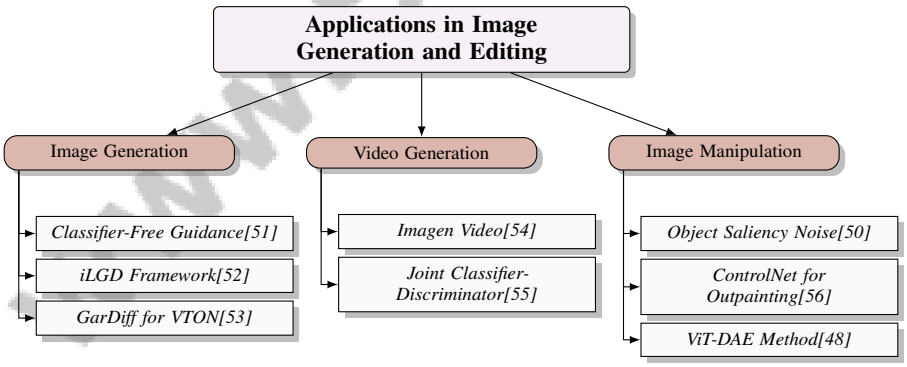


Figure 3: This figure illustrates the hierarchical categorization of applications in image generation and editing, highlighting key methods and innovations such as Classifier-Free Guidance, iLGD Framework, and GarDiff for virtual try-on in image generation; Imagen Video and Joint Classifier-Discriminator in video generation; and Object Saliency Noise, ControlNet for outpainting, and ViT-DAE Method in image manipulation.

4 Object Detection and Recognition

4.1 Techniques and Algorithms for Object Detection and Recognition

Object detection and recognition are pivotal in computer vision, focusing on identifying and classifying objects within images. Traditional methods like Histogram of Oriented Gradients (HOG) and

Scale-Invariant Feature Transform (SIFT) laid the groundwork for feature extraction but struggled with visual variations such as lighting, scale, and occlusion [36]. The advent of deep learning, particularly Convolutional Neural Networks (CNNs), revolutionized this field. Architectures such as Region-based Convolutional Neural Network (R-CNN) and its variants (Fast R-CNN, Faster R-CNN) enhanced accuracy and efficiency by leveraging deep hierarchical feature representations [59].

Recent advances have seen the integration of diffusion models and symbolic reasoning. The DeepMask method, for instance, improves detection and segmentation by using a CNN to predict segmentation masks and object likelihood scores simultaneously [60]. The Context-Dependent Diffusion Network (CDDN) models visual relationships among objects as subject-predicate-object triplets, enhancing relational understanding in complex scenes [61].

Diffusion models have also been employed in anomaly detection frameworks like LafitE, which utilize latent diffusion models to learn normal feature distributions and perform feature editing for improved anomaly detection [62]. The Diffusion-based Data Generator (DUR) converts Gaussian noise into high-fidelity images conditioned on class and distance, thus enhancing training data generation for object detection tasks [63].

Innovative methods like DODA generate high-quality detection data using layout images and reference images from the target domain, addressing data scarcity and annotation labor challenges [64]. The Hybrid Optimized Deep Convolutional Neural Network (HODCNN) further elevates detection accuracy by integrating advanced pre-processing and segmentation techniques [65].

These advancements not only enhance detection and recognition systems' accuracy and robustness but also expand their applicability across domains with complex visual conditions. Continuous evolution in these techniques promises further improvements, driven by integrating innovative models and methodologies, categorized into fundamental studies, skill-centric planning, safety mechanisms, and domain-specific applications [34].

4.2 Integration of Diffusion Models in Object Detection

The integration of diffusion models into object detection frameworks has significantly enhanced detection capabilities through robust feature extraction and iterative refinement. As illustrated in Figure 4, this figure highlights the integration of diffusion models into object detection, showcasing enhanced detection methods, innovative frameworks, and advanced techniques. Key contributions such as Latent Diffusion Models (LDMs), the Mamba Framework, and Transformer Backbones are emphasized, demonstrating their impact on improving object detection capabilities [22].

LDMs enable efficient image synthesis while preserving semantic details crucial for training detection algorithms [22]. Denoising Diffusion Probabilistic Models (DDPMs) generate high-quality synthetic images that bolster target recognition in complex environments [6].

Style Extracting Diffusion Models (STEDM) improve object detection by producing diverse images that reflect unseen styles, thus enhancing segmentation accuracy [66]. The Mamba framework's spatial and frequency scanning mechanisms allow for better representation of local and global features, surpassing existing models [67].

The Diffusion-based Data Generator (DUR) transforms Gaussian noise into high-fidelity images conditioned on class and distance, addressing data scarcity and annotation labor in object detection [45]. Diffusion models' adaptability is evident in medical imaging applications, generating realistic surgical videos [32], and reducing object expansion in salient object-aware background generation [56].

Moreover, integrating diffusion models with transformer backbones simplifies the incorporation of text and image features, enhancing object detection processes [26]. The CAAT method employs gradient-based optimization to apply subtle perturbations, improving adversarial attacks on customized diffusion models [68].

These advancements tackle challenges like spatial feature aliasing and computational expense, optimizing diffusion models through innovative reinforcement learning techniques. As diffusion models evolve, they are set to significantly enhance object detection capabilities across diverse applications, utilizing sophisticated image data augmentation techniques to generate high-fidelity, diverse images that strengthen training datasets. Their ability to capture discriminative features during

the generative process boosts performance in classification tasks, making them a powerful asset in computer vision. Emerging retrieval-augmented architectures illustrate how integrating external image databases can optimize generative models for effective zero-shot learning and adaptation to new domains without extensive retraining [69, 14, 18, 17].

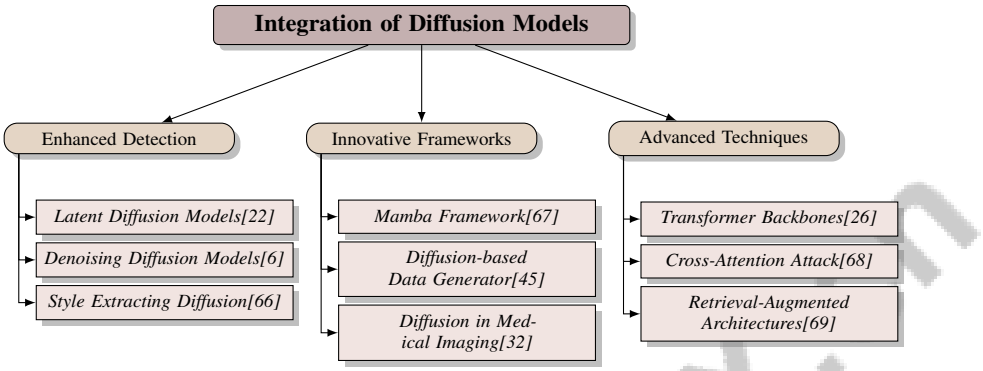


Figure 4: This figure illustrates the integration of diffusion models into object detection, highlighting enhanced detection methods, innovative frameworks, and advanced techniques. It includes key contributions like Latent Diffusion Models, the Mamba Framework, and Transformer Backbones, showcasing their impact on improving object detection capabilities.

4.3 Diffusion Models in Object Recognition

Diffusion models have significantly advanced object recognition by enhancing the generation of comprehensive semantic representations and improving recognition capabilities across various contexts. The DiffusionInst framework exemplifies this by formulating instance segmentation as a noise-to-filter denoising process, representing instances as instance-aware filters, thereby refining object recognition tasks and facilitating accurate outcomes [70].

In Ultra-Range Gesture Recognition (URGR) frameworks, Diffusion in Ultra-Range (DUR) generates synthetic images of objects at varying distances, crucial for training models in gesture recognition tasks, thereby enhancing performance in diverse operational scenarios [63]. Diffusion models also mitigate human effort in data labeling and scale the generation of diverse datasets, essential for improving model training and recognition accuracy [9].

Diffusion models excel in addressing inter-concept confusion, particularly in generating customized concepts. The CLIF methodology enhances the generation of concepts in Text-to-Image Generative Diffusion Models (TGDMs), effectively preserving the unique identities of concepts amidst complex interactions [71]. This capability improves the model's ability to maintain distinct semantic attributes, enhancing object recognition in scenarios with multiple interacting objects.

In few-shot settings, the DRDM framework leverages semantic relationships and external knowledge to enhance generalization and reduce feature contamination, boosting recognition accuracy in data-scarce environments [72]. The iterative optimization approach proposed by Zafar et al. improves the counting ability of diffusion models by optimizing generated images based on a counting loss derived from a counting model [73].

The TrackDiffusion method provides fine-grained control over object trajectories, ensuring high temporal consistency across generated video frames [74]. Furthermore, Rahman et al. showcase significant improvements in generating images with multiple interacting subjects, advancing personalized image generation in generative AI [58]. This development underscores the transformative impact of diffusion models on object recognition, paving the way for more accurate, efficient, and versatile recognition systems across a wide range of applications.

5 Image Segmentation Techniques

5.1 Traditional and Deep Learning-Based Segmentation Methods

Traditional segmentation methods, such as regional active contours, delineate object boundaries using edge detection and regional homogeneity but struggle with complex images due to intensity inhomogeneity and noise [75]. The advent of deep learning, particularly CNNs, has significantly improved segmentation accuracy and robustness across tasks like semantic, instance, panoptic, video, and 3D segmentation [38]. These models leverage large datasets and computational power to learn intricate feature representations.

As illustrated in Figure 5, the categorization of segmentation methods into traditional, deep learning, and diffusion model-based techniques highlights key advancements and specific frameworks within each category. Recent diffusion model advancements have further enhanced segmentation. The DDPS framework uses denoising diffusion with mask priors for improved semantic segmentation [76]. The LD-ZNet utilizes LDMs' rich semantic features for effective segmentation of natural and AI-generated images [77]. MOFT extracts motion information from video diffusion features, enabling motion control in segmentation [78]. The Diffuse, Attend, and Segment framework uses attention maps from stable diffusion models to enhance segmentation accuracy [79].

Synthetic datasets, generated under varied conditions, are essential for testing segmentation robustness, offering diverse scenarios for performance evaluation [80]. Example-based sampling improves efficiency in generating diverse point sets, maintaining sample properties while enhancing segmentation efficiency [81]. The transition from traditional to deep learning and diffusion-based techniques marks significant progress in image segmentation, applicable in industries like healthcare, transportation, and robotics. Innovations in generative AI, such as LDMs, leverage text-based inputs to enhance semantic understanding of image boundaries, improving pixel-level scene understanding and generating diverse, context-aware images [38, 77, 14].

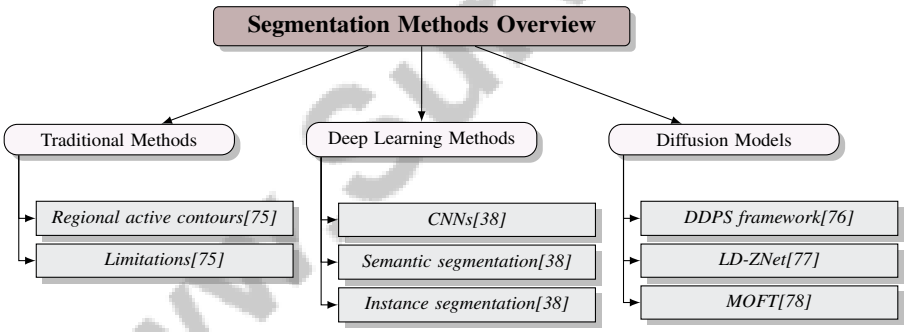


Figure 5: This figure illustrates the categorization of segmentation methods into traditional, deep learning, and diffusion model-based techniques, highlighting key advancements and specific frameworks within each category.

5.2 Integration of Diffusion Models in Image Segmentation

Integrating diffusion models into image segmentation has enhanced accuracy and efficiency through advanced methodologies. DiffSegmenter uses attention mechanisms to establish semantic relationships between text and visual features, improving segmentation precision [82]. In medical imaging, Temporal Reasoning Modules (TRM) in video polyp segmentation capture dynamic cues for enhanced accuracy [83]. DiffDIS introduces a one-step denoising process for detailed segmentation masks from high-resolution images, enhancing task efficiency [33]. The SDXL-Turbo diffusion model improves image generation accuracy in generative dataset distillation, crucial for effective segmentation [45].

Exploring temporal dynamics within diffusion models reveals redundancies and optimizes inference processes, enhancing computational efficiency [46]. Diffusion models improve salient object representation, showcasing versatility in handling complex background and foreground interactions [56]. These advancements illustrate diffusion models' transformative impact on image segmentation, enhancing outcomes in computer vision tasks through effective image data augmentation. Latent

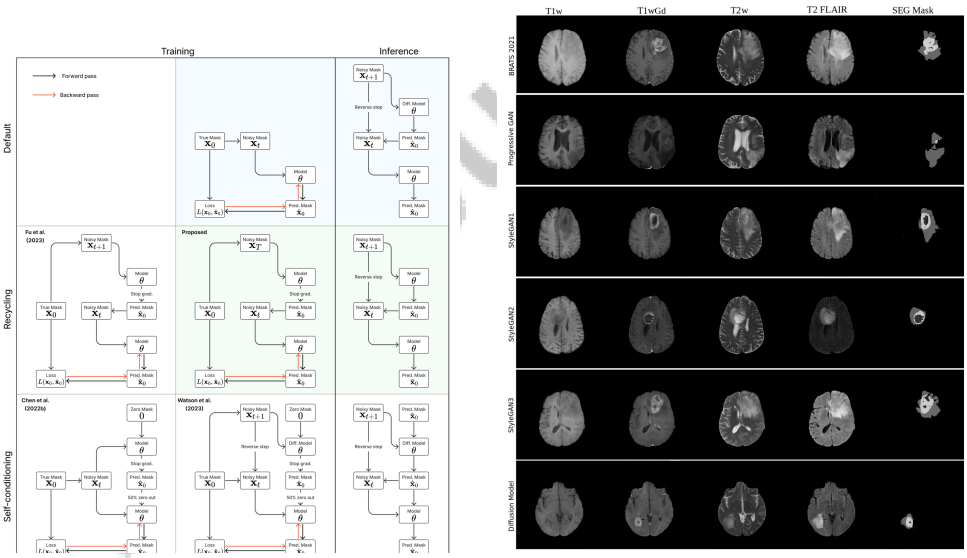
diffusion models (LDMs) excel in in-context segmentation, achieving competitive results against specialized models through innovative output alignment and optimization strategies [12, 14].

5.3 Applications in Medical Imaging and Other Domains

Diffusion models have significantly advanced medical imaging segmentation, addressing traditional method challenges. The MedSegDiff-V2 framework outperforms state-of-the-art methods across 20 medical image segmentation tasks, setting a new benchmark [40]. Fu et al.'s recycling training strategy improves segmentation performance across multiple datasets, demonstrating diffusion models' adaptability [41]. Akbar et al. highlight the utility of synthetic medical images for training, achieving performance comparable to real-image-trained models [84].

Pinaya et al. compiled a diverse dataset, including 2D chest X-ray images, 2D mammograms, 3D brain MRI, and 2D OCT images, ensuring broad applicability of diffusion models in medical imaging [85]. Beyond medical imaging, diffusion models are effective in other domains. The SSCC method, evaluated on the LIDC dataset, exemplifies segmentation techniques in medical imaging, particularly in chest CT scans [86]. This underscores segmentation's importance in accurately delineating complex anatomical structures for diagnosis and treatment planning.

The HiDiff framework enhances medical image segmentation by integrating discriminative and generative approaches [87]. Score-based generative models using Signed Distance Functions (SDFs) generate smoother segmentation masks, improving quality [88]. Diffusion models also excel in high-precision dichotomous image segmentation, as shown by the DiffDIS framework [33]. This capability extends diffusion models' applicability to domains requiring precise segmentation, advancing theoretical and practical aspects of image segmentation.



(a) Comparison of Different Masking Strategies in Deep Learning Models[41]

(b) Comparison of MRI Images and Segmentation Masks[84]

Figure 6: Examples of Applications in Medical Imaging and Other Domains

As shown in Figure 6, image segmentation is pivotal in various domains, particularly medical imaging, where it facilitates precise identification and delineation of structures within complex image data. The first part of the example showcases different masking strategies in deep learning models, illustrating how methodologies, including "Default," "Recycling," and approaches by Chen et al. (2022b) and Watson et al. (2023), influence data and gradient flow during training and inference phases. This comparison is crucial for understanding the impact of various strategies on segmentation performance. The second part presents a comparison of MRI images and their corresponding segmentation masks in a 3x3 grid format, allowing analysis of different MRI scan types (T1w, T1wGd, T2w, T2 FLAIR) alongside their segmentation outcomes, depicted in grayscale to indicate intensity variations. Together,

these examples underscore the importance of image segmentation techniques in enhancing medical imaging analysis and their potential applicability in other fields [41, 84].

6 Generative Models and Their Applications

6.1 Integration with Other Generative Models

The fusion of diffusion models with other generative frameworks has notably improved synthesis quality and efficiency, offering enhanced control in the generative process. The Ditail framework exemplifies this by employing multiple diffusion models for content and style transfer, allowing for a versatile approach to image generation [89]. The Image Retrieval Score (IRS) serves as a robust evaluation metric, assessing diversity in image generation by determining how many real images can be retrieved using synthetic data [90]. The RDM framework, which incorporates nearest neighbor retrieval during training and inference, ensures that generated images are diverse and contextually relevant [18]. Pinaya et al. demonstrate the versatility of integrating diffusion models with various architectures, including GANs and autoregressive transformers, to develop robust generative frameworks [85].

In text-to-image generation, Rahman et al. propose a method using alternating optimization to learn custom tokens and latent masks, enhancing generative output specificity [58]. Zhao et al. introduce a correlation-based approach to maintain the integrity of generated images by ensuring uncorrelated representations of omitted concepts [91]. Theoretical advancements have expanded understanding of generative diffusion models, with studies deriving theoretical bounds on total variation distance for mixed training scenarios [4]. Furthermore, exploring transformer backbones in diffusion models has yielded promising results, achieving a competitive FID score [26].

The integration of diffusion models with other generative approaches is set to significantly enhance synthesis quality and efficiency. Innovations such as retrieval-augmented diffusion models have improved synthesis quality by leveraging external image databases, facilitating tasks like zero-shot stylization and text-to-image synthesis [19, 92, 18, 10, 90]. The transition from 2D to 3D generative models indicates ongoing efforts to develop efficient representations for complex shapes and appearances, paving the way for practical applications in fields such as film and audiovisual arts.

6.2 Applications Beyond Traditional Domains

Generative models, particularly those enhanced by diffusion processes, have extended their applications beyond traditional domains like text-to-image generation into innovative interdisciplinary fields. They enable text-guided creative generation, producing diverse and contextually rich visual content from textual inputs, thus enhancing creative workflows while raising ethical considerations [19].

In medical imaging, diffusion models exhibit potential beyond conventional applications, such as fast unsupervised brain anomaly detection, indicating their adaptability for broader medical imaging contexts [93]. Their versatility is further evidenced in complex and dynamic domains, where advanced generative frameworks facilitate the automated creation of high-quality visual data for simulation environments in autonomous driving systems. By generating realistic scenarios with accurate object representations and contextual backgrounds, these models enhance testing and development processes [9, 94, 92, 54].

The integration of diffusion models into generative frameworks has broadened their applicability, enabling advancements in image synthesis through novel architectures that enhance visual quality. This expansion is attributed to the scalability and increased complexity of these models, alongside the strategic use of retrieval-augmented techniques for efficient training on diverse tasks. Recent research highlights the potential of diffusion models in handling structured data, such as tabular and time series information, achieving competitive performance in areas like zero-shot stylization and text-to-image synthesis [95, 18, 2]. As research continues to explore and refine these applications, diffusion models are poised to drive innovation across multiple fields, offering new possibilities and addressing complex challenges in both theoretical and practical dimensions.

7 Interconnections and Synergies

The exploration of methodologies in computer vision underscores the significant role of innovative frameworks and techniques in advancing generative models, particularly diffusion models, by enhancing image synthesis quality and integrating technologies like text-to-image frameworks. Retrieval-augmented diffusion models exemplify this by producing high-quality images with reduced complexity and computational demands. Furthermore, the generation of large-scale labeled training data via text-to-image synthesis optimizes training processes and enhances the robustness of machine learning models in tasks such as object detection and segmentation. These developments facilitate the versatile application of generative AI across various domains [9, 14, 18]. This section delves into innovative frameworks and techniques that leverage synergies among diffusion models, object detection, recognition, and image segmentation to drive progress in the field.

7.1 Innovative Frameworks and Techniques

In computer vision, innovative frameworks increasingly exploit synergies among diffusion models, object detection, recognition, and image segmentation to enhance performance. The integration of diffusion models with transformer architectures has notably improved performance across various vision tasks, achieving competitive Fréchet Inception Distance (FID) scores indicative of high-quality image synthesis [26]. DiffusionInst, for example, demonstrates the potential of diffusion models in refining object recognition and segmentation by framing instance segmentation as a noise-to-filter denoising process, capturing intricate details for accurate recognition [70].

The RDM framework further illustrates the integration of diffusion models with retrieval mechanisms, enhancing generative processes to ensure that generated images are diverse and contextually relevant [18]. In medical imaging, the MedSegDiff-V2 framework sets a new standard for segmentation tasks, outperforming existing state-of-the-art methods across multiple datasets [40]. This exemplifies how diffusion models integrated with advanced segmentation techniques can enhance accuracy and efficiency in complex scenarios.

In video generation, the TrackDiffusion method demonstrates fine-grained control over object trajectories while maintaining temporal consistency across frames [74], broadening the applicability of diffusion models in video-based recognition tasks. These frameworks and techniques underscore the transformative potential of leveraging synergies among diffusion models, object detection, recognition, and image segmentation. By integrating advanced methodologies, including text-to-image synthesis for training data generation, researchers can significantly enhance outcomes in computer vision, enabling the creation of accurately labeled synthetic datasets that rival real data performance [9, 37].

7.2 Synergies in Semantic Segmentation

The integration of diffusion models into semantic segmentation frameworks has markedly improved segmentation accuracy and efficiency, addressing persistent challenges in the field. Diffusion models, known for their iterative refinement processes, provide a robust mechanism for generating high-fidelity data representations essential for precise segmentation outcomes [39]. A key advantage is their effectiveness in utilizing limited labeled data, enabling high-quality segmentations with minimal supervision [39], which is particularly beneficial in medical imaging where labeled data is scarce.

Diffusion models also exhibit adaptability in incorporating various conditioning modalities, such as segmentation maps and sketches, into pretrained models without extensive retraining, enhancing their ability to capture complex semantic relationships within images [29]. In medical imaging, the MedSegDiff-V2 framework has set new benchmarks by outperforming existing state-of-the-art methods across multiple tasks, illustrating how diffusion models can leverage denoising capabilities for clearer, more precise segmentations [40].

The integration of diffusion models with attention mechanisms, as seen in the Diffuse, Attend, and Segment framework, further enhances segmentation accuracy by aggregating attention maps derived from stable diffusion models [79]. This approach underscores the significance of attention mechanisms in improving semantic understanding, leading to more accurate segmentation outcomes. The synergy between diffusion models and semantic segmentation frameworks has led to substantial advancements in computer vision, particularly through improved diversity and quality of train-

ing datasets via effective image data augmentation. This synergy enhances the performance and robustness of machine learning models, enabling nuanced and context-aware image editing.

Recent developments, including the application of information-theoretic principles to analyze visual-semantic relationships within diffusion models, further elucidate the mechanisms behind their success. Innovative approaches like Diffusion Feature Fusion (DIFF) have demonstrated the capacity to extract and integrate robust semantic representations, significantly enhancing cross-domain semantic segmentation and achieving state-of-the-art benchmarks. These advancements pave the way for more precise interpretations of visual data, thereby enhancing the capabilities of computer vision applications [12, 43, 14, 15]. By leveraging the strengths of diffusion models, researchers can achieve enhanced segmentation outcomes, driving innovation in both theoretical and applied aspects of computer vision.

8 Challenges and Future Directions

8.1 Challenges and Solutions

The application of diffusion models in computer vision encounters several obstacles, notably in computational efficiency, data quality, and model robustness. The iterative processes inherent to diffusion models demand significant computational resources, limiting their broader adoption [22]. Although Latent Diffusion Models (LDMs) aim to alleviate these demands, the substantial resources required for training and inference remain a critical concern [3]. The inefficiency in training large diffusion models, coupled with the need for extensive computational resources, complicates their deployment [44].

Another issue is the dependence on the quality of style query images, which can constrain the effectiveness of generated outputs [66]. In medical imaging, the scarcity of high-quality training data negatively impacts generative model performance [32]. Additionally, current models struggle with capturing global and long-range image relations due to reliance on spatial processing and fixed scanning orders, limiting their ability to fully utilize the frequency spectrum [67].

Generalization across tasks and datasets also poses challenges. Overfitting risks with small datasets can impede generalization, especially in high-precision image segmentation [33]. Moreover, existing methods often fail to exploit the adaptable structure of Unets and the importance of time steps, leading to suboptimal inference performance [46]. The loss of crucial localization information and shape awareness in current discriminative models hampers effective segmentation, as highlighted by DiffSegmenter [82]. Furthermore, existing benchmarks lack comprehensive evaluation criteria for assessing unlearning performance across diverse scenarios, such as style and object unlearning [47].

To address these challenges, several solutions have been proposed. Developing more efficient training algorithms can reduce computational demands, enhancing the accessibility of diffusion models [44]. Integrating adaptive pipelines with existing architectures may mitigate issues related to non-optimal fits. Adapting diffusion models for inpainting methods can also help address challenges in modifying salient object boundaries [56]. Furthermore, exploring the potential of combining multiple diffusion models, rather than focusing solely on individual improvements, could unlock novel image generation opportunities and enhance synthesis outcomes. Implementing these solutions could effectively mitigate the challenges associated with diffusion models, fostering broader adoption and innovation in the field.

8.2 Future Directions

The future of diffusion models in computer vision is set to explore several promising directions aimed at enhancing efficiency, applicability, and integration with other technologies. A critical focus will be on improving computational efficiency and real-time decision-making capabilities, essential for refining human-robot interactions through advanced models [34]. Optimizing training processes and investigating novel noise distributions could significantly enhance sampling efficiency, expanding applicability across diverse domains. Architectural improvements and novel image transforms to reformulate the DDPM objective and noising schedule will also be crucial for enhancing out-of-distribution detection capabilities.

Future research could involve developing more effective generative dataset distillation techniques that adapt to various datasets, improving the quality and efficiency of generated images [45]. Additionally, exploring the implications of content replication on copyright, with a focus on robust methods for detecting and mitigating replication in generative models, is vital [6].

In model architectures, future work could focus on optimizing intervention strategies and exploring their applicability across different architectures [46]. Expanding machine unlearning applications to various model types and scenarios represents another promising avenue [47].

Exploring background generation for non-salient objects and investigating alternative control architectures could further enhance diffusion model capabilities [56]. Additionally, optimizing methods to mitigate societal impacts of generated deepfake videos is an important future research direction [96].

In medical imaging, prioritizing diverse applications for diffusion models, particularly through Fourier Diffusion Models, can enhance image quality and measurement accuracy. These models address limitations in existing methods by enabling better control over modulation transfer function (MTF) and noise power spectrum (NPS) during image generation. Innovations like Temporal Harmonization for Optimal Restoration (THOR) show potential for improving clinical applications by refining anomaly detection and segmentation in medical scans, preserving healthy tissue integrity and reducing false positive rates [97, 23, 32, 2]. Enhancing generalization capabilities, generating segmentation targets, and creating temporally connected data will be crucial for advancing medical imaging techniques. Exploring multi-class segmentation and the joint distribution of images and segmentation masks could further extend the applicability of diffusion models in this domain.

Furthermore, advancements in garment representation and real-time virtual try-on systems, as well as extending diffusion transformers to more complex dynamic models, are promising areas for future exploration. In adversarial robustness, enhancing methods such as CAAT to address complex adversarial scenarios and their applicability to a broader range of generative models is essential. Understanding how subtle perturbations exploit vulnerabilities in various architectures, including latent diffusion models, and developing strategies to mitigate risks associated with generating synthetic media is crucial. Additionally, exploring CAAT's potential in improving model security across different generative frameworks will be vital for advancing the field [19, 80, 49, 18, 68].

By focusing on these identified research directions, diffusion models are set to significantly advance innovation in computer vision and artificial intelligence. Demonstrating exceptional capabilities in image data augmentation, text-to-image synthesis, and video generation, these models are expanding their influence across diverse applications, including image editing, semantic manipulation, and personalized content creation. Addressing current challenges and exploring future avenues will likely enhance the performance and robustness of machine learning systems, broadening their impact across various fields [20, 14, 19].

9 Conclusion

This survey delves into the pivotal contributions of diffusion models to computer vision, highlighting their transformative impact on object detection, recognition, image segmentation, and generative modeling. These models excel in generating high-fidelity data representations, as evidenced by their outstanding performance in realistic image synthesis and video generation. Their capacity to blend visual concepts highlights their adaptability, showcasing how different blending methods and concepts can influence outcomes.

In medical imaging, diffusion models have set new standards, particularly through frameworks like Diff-VPS, which establish essential baselines for video segmentation. The MONAI Generative Models benchmark enhances model evaluation, providing a reproducible and modular framework that significantly advances the field.

Despite their strengths, diffusion models face challenges related to computational efficiency and versatility across diverse data types. The reproducibility of these models is sensitive to training data size and model capacity, underscoring the importance of meticulous model design and data curation. Additionally, the phase transition observed in class representation within diffusion models reveals their complexity and suggests further exploration into class dynamics.

Recent advancements, such as bridging discrete and continuous modeling, have shown promising results, advancing the development of unified diffusion models. Furthermore, assessing the influence of training data on diffusion model outputs has offered valuable insights into model behavior and data dependencies.

References

[1] Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers, 2024.

[2] Melike Nur Yeğin and Mehmet Fatih Amasyalı. Theoretical research on generative diffusion models: an overview, 2024.

[3] Harshith Bachimanchi and Giovanni Volpe. Diffusion models to enhance the resolution of microscopy images: A tutorial, 2024.

[4] Shi Fu, Sen Zhang, Yingjie Wang, Xinmei Tian, and Dacheng Tao. Towards theoretical understandings of self-consuming generative models, 2024.

[5] Huijie Zhang, Yifu Lu, Ismail Alkhouri, Saiprasad Ravishankar, Dogyoon Song, and Qing Qu. Improving efficiency of diffusion models via multi-stage framework and tailored multi-decoder architectures, 2024.

[6] Diffusion art or digital forgery? investigating data replication in diffusion models.

[7] Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris Metaxas, and Ahmed Elgammal. Diffusion guided domain adaptation of image generators, 2022.

[8] Hengyu Fu, Zehao Dou, Jiawei Guo, Mengdi Wang, and Minshuo Chen. Diffusion transformer captures spatial-temporal dependencies: A theory for gaussian process data, 2025.

[9] Yunhao Ge, Jiashu Xu, Brian Nlong Zhao, Neel Joshi, Laurent Itti, and Vibhav Vineet. Beyond generation: Harnessing text to image models for object detection and segmentation, 2023.

[10] Luís Arandas, Mick Grierson, and Miguel Carvalhais. Antagonising explanation and revealing bias directly through sequencing and multimodal inference, 2023.

[11] E. Zhixuan Zeng, Yuhao Chen, and Alexander Wong. Understanding the limitations of diffusion concept algebra through food, 2024.

[12] Chaoyang Wang, Xiangtai Li, Henghui Ding, Lu Qi, Jiangning Zhang, Yunhai Tong, Chen Change Loy, and Shuicheng Yan. Explore in-context segmentation via latent diffusion models, 2024.

[13] Lorenzo Olearo, Giorgio Longari, Simone Melzi, Alessandro Raganato, and Rafael Peñaloza. How to blend concepts in diffusion models, 2024.

[14] Panagiotis Alimisis, Ioannis Mademlis, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, and Georgios Th. Papadopoulos. Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions, 2025.

[15] Rushikesh Zawar, Shaurya Dewan, Prakanshul Saxena, Yingshan Chang, Andrew Luo, and Yonatan Bisk. Diffusionpid: Interpreting diffusion via partial information decomposition, 2024.

[16] Clinton J. Wang and Polina Golland. Interpolating between images with diffusion models, 2023.

[17] Raghav Singhal, Mark Goldstein, and Rajesh Ranganath. Where to diffuse, how to diffuse, and how to get back: Automated learning for multivariate diffusions, 2023.

[18] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022.

[19] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, In So Kweon, and Junmo Kim. Text-to-image diffusion models in generative ai: A survey, 2024.

[20] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42, 2024.

[21] Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data, 2024.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[23] Matthew Tivnan, Jacopo Teneggi, Tzu-Cheng Lee, Ruoqiao Zhang, Kirsten Boedeker, Liang Cai, Grace J. Gang, Jeremias Sulam, and J. Webster Stayman. Fourier diffusion models: A method to control mtf and nps in score-based stochastic image generation, 2023.

[24] Unifying diffusion models' latent space with applications to cyclediffusion and guidance.

[25] Yuewei Yang, Jialiang Wang, Xiaoliang Dai, Peizhao Zhang, and Hongbo Zhang. An analysis on quantizing diffusion transformers, 2024.

[26] Princy Chahal. Exploring transformer backbones for image diffusion models, 2022.

[27] Tom Fischer, Pascal Peter, Joachim Weickert, and Eddy Ilg. Neuroexplicit diffusion models for inpainting of optical flow fields, 2024.

[28] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua 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*, pages 1206–1217, 2023.

[29] Cusuh Ham, James Hays, Jingwan Lu, Krishna Kumar Singh, Zhifei Zhang, and Tobias Hinz. Modulating pretrained diffusion models for multimodal image synthesis, 2023.

[30] Florence Regol and Mark Coates. Diffusing gaussian mixtures for generating categorical data, 2023.

[31] Zijian Zhang, Zhou Zhao, Jun Yu, and Qi Tian. Shiftddpms: Exploring conditional diffusion models by shifting diffusion trajectories, 2023.

[32] Xi Zhu, Wei Zhang, Yijie Li, Lauren J. O'Donnell, and Fan Zhang. When diffusion mri meets diffusion model: A novel deep generative model for diffusion mri generation, 2024.

[33] Qian Yu, Peng-Tao Jiang, Hao Zhang, Jinwei Chen, Bo Li, Lihe Zhang, and Huchuan Lu. High-precision dichotomous image segmentation via probing diffusion capacity, 2024.

[34] Toshihide Ubukata, Jialong Li, and Kenji Tei. Diffusion model for planning: A systematic literature review, 2024.

[35] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition, 2020.

[36] Ulrich Hillenbrand and Gerd Hirzinger. Probabilistic search for object segmentation and recognition, 2002.

[37] Seyed-Mahdi Khaligh-Razavi. What you need to know about the state-of-the-art computational models of object-vision: A tour through the models, 2014.

[38] Yuanbo Wang, Unaiza Ahsan, Hanyan Li, and Matthew Hagen. A comprehensive review of modern object segmentation approaches, 2023.

[39] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models, 2022.

[40] Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion based medical image segmentation with transformer, 2023.

[41] Yunguan Fu, Yiwen Li, Shaheer U Saeed, Matthew J Clarkson, and Yipeng Hu. A recycling training strategy for medical image segmentation with diffusion denoising models, 2023.

[42] Anwaar Ulhaq and Naveed Akhtar. Efficient diffusion models for vision: A survey, 2024.

[43] Yuxiang Ji, Boyong He, Chenyuan Qu, Zhuoyue Tan, Chuan Qin, and Liaoni Wu. Diffusion features to bridge domain gap for semantic segmentation, 2024.

[44] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers, 2024.

[45] Duo Su, Junjie Hou, Guang Li, Ren Togo, Rui Song, Takahiro Ogawa, and Miki Haseyama. Generative dataset distillation based on diffusion model, 2024.

[46] Vidya Prasad, Chen Zhu-Tian, Anna Vilanova, Hanspeter Pfister, Nicola Pezzotti, and Hendrik Strobelt. Unraveling the temporal dynamics of the unet in diffusion models, 2023.

[47] Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Rao Kompella, Xiaoming Liu, and Sijia Liu. Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models, 2024.

[48] Xuan Xu, Saarthak Kapse, Rajarsi Gupta, and Prateek Prasanna. Vit-dae: Transformer-driven diffusion autoencoder for histopathology image analysis, 2023.

[49] This iccv paper is the open acce.

[50] Vedant Singh, Surgan Jandial, Ayush Chopra, Siddharth Ramesh, Balaji Krishnamurthy, and Vineeth N. Balasubramanian. On conditioning the input noise for controlled image generation with diffusion models, 2022.

[51] Yannik Frisch, Moritz Fuchs, Antoine Sanner, Felix Anton Ucar, Marius Frenzel, Joana Wasielica-Poslednik, Adrian Gericke, Felix Mathias Wagner, Thomas Dratsch, and Anirban Mukhopadhyay. Synthesising rare cataract surgery samples with guided diffusion models, 2023.

[52] Zakaria Patel and Kirill Serkh. Enhancing image layout control with loss-guided diffusion models, 2024.

[53] Siqi Wan, Yehao Li, Jingwen Chen, Yingwei Pan, Ting Yao, Yang Cao, and Tao Mei. Improving virtual try-on with garment-focused diffusion models, 2024.

[54] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022.

[55] Shelly Golan, Roy Ganz, and Michael Elad. Enhancing consistency-based image generation via adversarialy-trained classification and energy-based discrimination, 2024.

[56] Amir Erfan Eshratifar, Joao V. B. Soares, Kapil Thadani, Shaunak Mishra, Mikhail Kuznetsov, Yueh-Ning Ku, and Paloma de Juan. Salient object-aware background generation using text-guided diffusion models, 2024.

[57] Theodoros Kouzelis, Manos Plitsis, Mihalis A. Nicolaou, and Yannis Panagakis. Enabling local editing in diffusion models by joint and individual component analysis, 2024.

[58] Tanzila Rahman, Shweta Mahajan, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Leonid Sigal. Visual concept-driven image generation with text-to-image diffusion model, 2024.

[59] Hao Wang, Xingyu Lin, Yimeng Zhang, and Tai Sing Lee. Learning robust object recognition using composed scenes from generative models, 2017.

[60] Pedro O. Pinheiro, Ronan Collobert, and Piotr Dollar. Learning to segment object candidates, 2015.

[61] Zhen Cui, Chunyan Xu, Wenming Zheng, and Jian Yang. Context-dependent diffusion network for visual relationship detection, 2018.

[62] Haonan Yin, Guanlong Jiao, Qianhui Wu, Borje F. Karlsson, Biqing Huang, and Chin Yew Lin. Lafite: Latent diffusion model with feature editing for unsupervised multi-class anomaly detection, 2023.

[63] Eran Bamani, Eden Nissinman, Lisa Koenigsberg, Inbar Meir, and Avishai Sintov. A diffusion-based data generator for training object recognition models in ultra-range distance, 2024.

[64] Shuai Xiang, Pieter M. Blok, James Burridge, Haozhou Wang, and Wei Guo. Doda: Diffusion for object-detection domain adaptation in agriculture, 2024.

[65] Venkata Beri. Hybrid optimized deep convolution neural network based learning model for object detection, 2022.

[66] Mathias Öttl, Frauke Wilm, Jana Steenpass, Jingna Qiu, Matthias Rübner, Arndt Hartmann, Matthias Beckmann, Peter Fasching, Andreas Maier, Ramona Erber, Bernhard Kainz, and Katharina Breininger. Style-extracting diffusion models for semi-supervised histopathology segmentation, 2024.

[67] Hao Phung, Quan Dao, Trung Dao, Hoang Phan, Dimitris Metaxas, and Anh Tran. Dimsum: Diffusion mamba – a scalable and unified spatial-frequency method for image generation, 2025.

[68] Jingyao Xu, Yuetong Lu, Yandong Li, Siyang Lu, Dongdong Wang, and Xiang Wei. Perturbing attention gives you more bang for the buck: Subtle imaging perturbations that efficiently fool customized diffusion models, 2024.

[69] Soumik Mukhopadhyay, Matthew Gwilliam, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Srinidhi Hegde, Tianyi Zhou, and Abhinav Shrivastava. Diffusion models beat gans on image classification, 2023.

[70] Diffusioninst: Diffusion model f.

[71] Wang Lin, Jingyuan Chen, Jiaxin Shi, Yichen Zhu, Chen Liang, Junzhong Miao, Tao Jin, Zhou Zhao, Fei Wu, Shuicheng Yan, and Hanwang Zhang. Non-confusing generation of customized concepts in diffusion models, 2024.

[72] Tianxu Wu, Shuo Ye, Shuhuang Chen, Qinmu Peng, and Xinge You. Detail reinforcement diffusion model: Augmentation fine-grained visual categorization in few-shot conditions, 2024.

[73] Oz Zafar, Lior Wolf, and Idan Schwartz. Iterative object count optimization for text-to-image diffusion models, 2024.

[74] Pengxiang Li, Kai Chen, Zhili Liu, Ruiyuan Gao, Lanqing Hong, Guo Zhou, Hua Yao, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Trackdiffusion: Tracklet-conditioned video generation via diffusion models, 2024.

[75] M. Abdelsamea. Regional active contours based on variational level sets and machine learning for image segmentation, 2015.

[76] Zeqiang Lai, Yuchen Duan, Jifeng Dai, Ziheng Li, Ying Fu, Hongsheng Li, Yu Qiao, and Wenhai Wang. Denoising diffusion semantic segmentation with mask prior modeling, 2023.

[77] Koutilya Pnvr, Bharat Singh, Pallabi Ghosh, Behjat Siddiquie, and David Jacobs. Ld-znet: A latent diffusion approach for text-based image segmentation, 2023.

[78] Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller, 2024.

[79] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion, 2024.

[80] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models, 2022.

[81] Bastien Doignies, Nicolas Bonneel, David Coeurjolly, Julie Digne, Loïs Paulin, Jean-Claude Iehl, and Victor Ostromoukhov. Example-based sampling with diffusion models, 2023.

[82] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter, 2024.

[83] Yingling Lu, Yijun Yang, Zhaohu Xing, Qiong Wang, and Lei Zhu. Diff-vps: Video polyp segmentation via a multi-task diffusion network with adversarial temporal reasoning, 2024.

[84] Muhammad Usman Akbar, Måns Larsson, and Anders Eklund. Brain tumor segmentation using synthetic mr images – a comparison of gans and diffusion models, 2024.

[85] Walter H. L. Pinaya, Mark S. Graham, Eric Kerfoot, Petru-Daniel Tudosiu, Jessica Dafflon, Virginia Fernandez, Pedro Sanchez, Julia Wolleb, Pedro F. da Costa, Ashay Patel, Hyungjin Chung, Can Zhao, Wei Peng, Zelong Liu, Xueyan Mei, Oeslle Lucena, Jong Chul Ye, Sotirios A. Tsaftaris, Prerna Dogra, Andrew Feng, Marc Modat, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Generative ai for medical imaging: extending the monai framework, 2023.

[86] Lukas Zbinden, Lars Doorenbos, Theodoros Pissas, Adrian Thomas Huber, Raphael Sznitman, and Pablo Márquez-Neila. Stochastic segmentation with conditional categorical diffusion models, 2023.

[87] Tao Chen, Chenhui Wang, Zhihao Chen, Yiming Lei, and Hongming Shan. Hidiff: Hybrid diffusion framework for medical image segmentation, 2024.

[88] Lea Bogensperger, Dominik Narnhofer, Filip Ilic, and Thomas Pock. Score-based generative models for medical image segmentation using signed distance functions, 2023.

[89] Haoming Liu, Yuanhe Guo, Shengjie Wang, and Hongyi Wen. Diffusion cocktail: Mixing domain-specific diffusion models for diversified image generations, 2024.

[90] Mischa Dombrowski, Weitong Zhang, Sarah Cechnicka, Hadrien Reynaud, and Bernhard Kainz. Image generation diversity issues and how to tame them, 2024.

[91] Mengnan Zhao, Lihe Zhang, Tianhang Zheng, Yuqiu Kong, and Baocai Yin. Separable multi-concept erasure from diffusion models, 2024.

[92] Zifan Shi, Sida Peng, Yinghao Xu, Andreas Geiger, Yiyi Liao, and Yujun Shen. Deep generative models on 3d representations: A survey, 2023.

[93] Walter H. L. Pinaya, Mark S. Graham, Robert Gray, Pedro F Da Costa, Petru-Daniel Tudosiu, Paul Wright, Yee H. Mah, Andrew D. MacKinnon, James T. Teo, Rolf Jager, David Werring, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Fast unsupervised brain anomaly detection and segmentation with diffusion models, 2022.

[94] Thomas Lips and Francis wyffels. Evaluating text-to-image diffusion models for texturing synthetic data, 2024.

[95] Heejoon Koo and To Eun Kim. A comprehensive survey on generative diffusion models for structured data, 2023.

[96] Vidm: Video implicit diffusion models.

[97] Cosmin I. Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A. Schnabel. Diffusion models with implicit guidance for medical anomaly detection, 2024.

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.