Artificial Intelligence Review (2026) 59:45
https://doi.org/10.1007/s10462-025-11450-0

Revisiting U-Net: a foundational backbone for modern generative AI

Marvin John Ignacio[1] · Sangyun Shin[2] · Hulin Jin[3] · Seong Joon Yoo[1] · Dongil Han[1] · Yong-Guk Kim[1]

Received: 1 May 2025 / Accepted: 11 November 2025
© The Author(s) 2025

Abstract
This survey explores the evolution and application of U-Net in generative AI, highlighting its success across various modalities, including image, text, audio, video, 3D, and pose/action generation. Initially designed for biomedical segmentation, U-Net has been adapted and enhanced with architectural innovations such as normalization techniques, self and cross-attention mechanisms, and residual connections. These advancements have made U-Net a powerful backbone for modern generative models in diffusion-based frameworks, GANs, and autoregressive architectures. The survey comprehensively reviews U-Net's modality-specific applications, from high-resolution image synthesis and text-to-image generation to speech enhancement, video generation, 3D reconstruction, and pose/action generation. Despite its widespread success, U-Net faces challenges in computational efficiency, contextual understanding, and scalability for multimodal tasks. Future directions focus on optimizing U-Net for lightweight and real-time applications, enhancing its contextual awareness, and improving its integration with emerging architectures like transformers and diffusion models.

Keywords U-Net · Generative AI · Modality · Diffusion models · GANs · Autoregressive models

1 Introduction

The rapid advancements in Generative AI have revolutionized how machines create and manipulate content, ranging from photorealistic images and videos to human-like speech and text. At the forefront of this transformation are powerful architectures, including Generative Adversarial Networks (GANs) (Iglesias et al. 2022), Variational Autoencoders (VAEs) (Asperti et al. 2021), Transformers (Lin et al. 2021), and Diffusion models (Cao et al. 2022). Transformer-based models have received substantial attention due to their success in large-scale language and vision tasks (Al-hammuri et al. 2023; Ansar et al. 2024; Han et al. 2020; Huang et al. 2023). Nevertheless, their adoption remains constrained by scalability issues, primarily due to the quadratic complexity of their attention mechanisms. While

Extended author information available on the last page of the article

Published online: 24 November 2025

Transformers continue to dominate recent generative modeling efforts, another foundational architecture, the U-Net, has become increasingly central to generative AI, particularly in diffusion-based frameworks. Although U-Net has been extensively studied in fields such as biomedical image segmentation, its broader contributions to generative AI across multiple modalities have not yet been surveyed in a systematic and comprehensive manner. This gap motivates the present study, which seeks to examine the architectural evolution of U-Net and to highlight its emerging role as a critical component in modern generative modeling.

Originally designed for biomedical segmentation (Ronneberger et al. 2015), U-Net has evolved into a versatile tool for generative modeling. This convolutional encoder–decoder architecture is renowned for its efficient feature extraction and high-resolution output reconstruction, characterized by its hallmark feature: *skip connections* that preserve spatial information throughout the network. While U-Net's initial application focused on segmentation, its architecture has proven to be particularly effective for generative tasks that require maintaining fine-grained details, such as image synthesis, denoising, inpainting, and even tasks beyond image generation, including text-to-image synthesis, speech enhancement, and 3D reconstruction.

What makes U-Net especially valuable in the context of generative AI is its flexibility across various data modalities. U-Net has been successfully integrated into generative frameworks for a wide range of applications, including **image generation**, **audio and speech synthesis**, **video generation**, **text modeling**, and even **3D volumetric reconstruction** and **human pose/action generation**. The versatility of U-Net is evident not only in its application across diverse domains but also in its integration within various model types, such as Diffusion models, GANs, and autoregressive models, where U-Net enhances the stability, feature propagation, and overall performance of these systems.

Despite the rapid expansion of generative AI research, existing survey papers remain largely model or application-centric (Cao et al. 2023; Gozalo-Brizuela and Garrido Merchan 2024; Qin and Hui 2023; Raut and Singh 2024; Zhou et al. 2024), emphasizing specific techniques such as Diffusion models, Large Language Models (LLMs), or Text-to-Image generation without analyzing the underlying architectural patterns that enable such advancements. In particular, the cross-domain significance of U-Net as a foundational backbone across diverse modalities has been largely overlooked. Meanwhile, surveys focusing on U-Net are typically constrained to traditional tasks such as biomedical image segmentation (Punn and Agarwal 2021; Wu et al. 2022) or audio enhancement (Gul and Khan 2023), failing to capture the broader evolution of U-Net as a crucial building block for modern generative models.

This gap highlights the need for a dedicated study that systematically explores how U-Net architectures have been adapted, extended, and integrated into cutting-edge generative AI frameworks. Our work addresses this gap by presenting the first comprehensive survey that situates U-Net within the context of generative AI across multiple modalities. Beyond simply cataloging applications, we highlight key architectural innovations that have shaped the evolution of U-Net. These include normalization strategies, attention mechanisms, and residual connections. Together, these advances have transformed U-Net from a basic encoder–decoder structure into a versatile and scalable backbone for Diffusion models, GANs, and Autoregressive generators. Furthermore, our work proposes new perspectives and future directions for incorporating U-Net-based designs into emerging generative frameworks, aiming to inspire further innovations in the field. Through this, we hope to

shed light on the often underappreciated, yet critical, role that U-Net plays in advancing the frontiers of generative AI research.

1.1 Contributions

This survey provides a comprehensive review of U-Net's role in generative modeling, highlighting its versatility across various data modalities. The key contributions of this work are as follows:

- A detailed review of the U-Net architecture, including its foundational taxonomy, structural components, and encoder–decoder methodology, establishing a systematic framework for understanding its generative design principles.
- A comprehensive analysis of U-Net's integration into Diffusion models, GANs, and autoregressive generative frameworks, clarifying its functional roles across different generative paradigms.
- A modality-centric overview of U-Net's applications in image synthesis, text and language modeling, audio and speech generation, video synthesis, 3D volumetric reconstruction, and human pose/action generation.
- A dedicated discussion of U-Net's limitations, highlighting its current architectural constraints in scalability, contextual modeling, and multimodal adaptation, which distinguishes them from broader research challenges.
- An evaluation of U-Net's architectural advantages, including its hierarchical feature extraction capabilities, fine-detail preservation through skip connections, and computational efficiency.
- A forward-looking analysis of current challenges and future directions, including the development of more efficient variants, Transformer fusion strategies, and multi-modal learning frameworks for large-scale generative AI.

In Sect. 2, the paper introduces a taxonomy of U-Net variants and outlines the core architectural design and review methodology used in this study. Section 3 then explores key architectural enhancements that have improved U-Net's performance in generative tasks, categorizing models based on their underlying training framework, including Diffusion, GAN, and Autoregressive paradigms. Section 4 presents a modality-centric analysis, examining how U-Net has been applied across various data types such as image, text, audio, video, 3D, and pose/action. Section 5 introduces a dedicated discussion of U-Net's current limitations, distinguishing architectural constraints from broader research challenges. Section 6 highlights the core advantages of U-Net across multiple domains, followed by Sect. 7, which outlines the open challenges that remain in its deployment. Section 8 then discusses emerging research opportunities and integration prospects for U-Net in next-generation generative AI systems. Finally, Sect. 9 concludes the paper by summarizing U-Net's overall impact and continuing evolution.

2 Taxonomy, architecture, and survey methodology

This section provides a foundational overview of U-Net's role in generative modeling, beginning with a taxonomy that organizes its diverse applications across generative paradigms and data modalities. This conceptual framework establishes the broader context for understanding how U-Net has evolved beyond its original purpose in biomedical image segmentation to become a core component of modern generative AI systems. Following the taxonomy, the discussion examines the architectural foundations of U-Net in detail, focusing on its encoder–decoder structure, skip connections, and their roles in enabling effective hierarchical feature extraction and reconstruction. The section concludes with an outline of the structured methodology used to identify and categorize relevant literature across multiple generative AI domains. Together, these components provide both the conceptual and technical basis necessary to understand U-Net's expanding capabilities and its central position within contemporary generative modeling research.

2.1 Taxonomy

The rapid evolution of U-Net from a biomedical segmentation model into a core component of modern generative systems has led to diverse adaptations across architectures and application domains. To provide a structured view of this landscape, Fig. 1 presents a taxonomy that organizes U-Net's roles according to two key axes: generative paradigms and data modalities. This taxonomy serves as a conceptual roadmap for understanding how U-Net supports a broad range of generative tasks and how its utility has expanded in parallel with advances in generative modeling.

At the first level, U-Net has been integrated into three principal generative paradigms: Generative Adversarial Networks (GANs), Diffusion-based models, and Autoregressive frameworks. Each of these paradigms leverages the U-Net's encoder–decoder structure and skip connections in distinct ways. In GAN-based systems, the U-Net is often employed as either a generator or a discriminator to enhance spatial detail preservation in tasks such as image-to-image translation and pose synthesis. In Diffusion models, the U-Net serves as the denoising backbone, iteratively refining noisy inputs into coherent outputs and enabling high-resolution synthesis with controllable generation. Autoregressive approaches, though less common, incorporate the U-Net as a sequential decoder that captures local spatial dependencies while maintaining a hierarchical structure.

At the second level, each generative paradigm further branches into data modalities, reflecting the U-Net's versatility across application domains. U-Net architectures have been specialized for image generation and translation, audio and speech synthesis, text-conditioned generation, video prediction, 3D volumetric reconstruction, and pose/action modeling. This modality-centric view highlights the U-Net's adaptability. While the fundamental encoder–decoder structure remains consistent, the architectural design and conditioning mechanisms are often tailored to modality-specific requirements, such as temporal coherence in video, cross-modal alignment in text-to-image generation, or spatial continuity in volumetric reconstruction.

Together, this taxonomy illustrates how U-Net's evolution is shaped not only by the underlying generative paradigm but also by the nature of the data and task. It provides a high-level framework for interpreting the subsequent sections of this survey, which examine

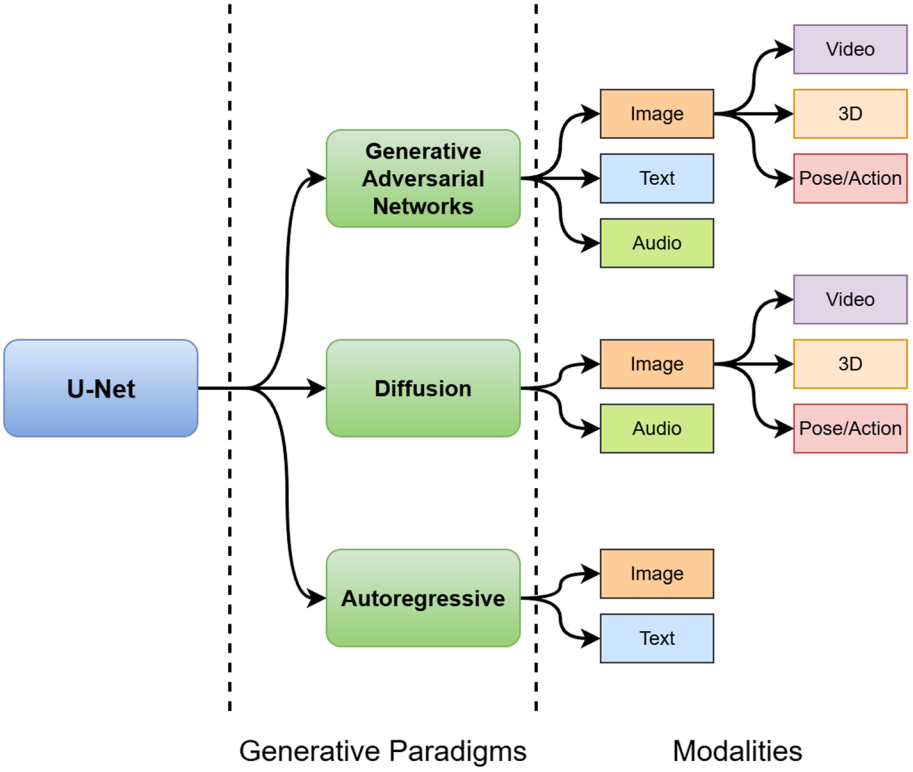

Fig. 1 Taxonomy of U-Net applications in generative AI, organized by paradigm and data modality. The figure illustrates how U-Net serves as a core architecture across generative adversarial networks, diffusion-based models, and autoregressive approaches, and highlights its adaptation to diverse data modalities, including image, text, audio, video, 3D, and pose or action generation

the architectural design of U-Net in detail and analyze how enhancements, such as attention mechanisms, normalization strategies, and hybrid backbones, further extend its generative capabilities.

2.2 Architectural structure

Let $\mathcal{X} \in \mathbb{R}^{D \times C}$ represent the input dataset, where each sample x_i has a dimension D corresponding to 1D, 2D, or 3D data. These input data types adhere to the conventions of convolutional operations, where kernels process adjacent features to extract meaningful patterns. For 1D data, the spatial or temporal length is denoted as L or T, respectively. Examples include token embeddings in natural language processing (spatial) and time series signals (temporal). In the case of 2D data, the most common form is an image with dimensions $H \times W$, where H represents the height and W represents the width. Extending to 3D data, input structures can be expressed as either $H \times W \times T$ for video sequences (incorporating temporal frames) or $H \times W \times L$ for volumetric data, such as CT scans (where L represents depth). Additionally, an extra feature dimension C encodes the information at each spatial, temporal, or volumetric location, such as color channels in images or embedding dimen-

sions in text representations. The structural overview of the standard U-Net architecture is shown in Fig. 2.

2.2.1 Encoder

The encoder, often termed the *contracting path*, is responsible for extracting hierarchical features from the input data through a series of downsampling operations. It is formally denoted as $\mathcal{E}_x$, where x is the input sample. The encoder consists of multiple sequential layers, each applying convolution, normalization, activation, and pooling operations. These transformations progressively reduce the spatial or temporal resolution of the input while increasing the number of feature channels, thereby enabling the model to learn increasingly abstract representations.

Given an input tensor $x_i \in \mathbb{R}^{D \times C_j}$, where C_j denotes the number of input channels, the first operation involves convolution. This process applies a learnable weight matrix, referred to as a kernel, which extracts localized patterns from the input. The kernel has a dimension of $K \times C_j \times C_{j+1}$, where K denotes the kernel size, C_j is the number of input channels, and C_{j+1} is the number of output channels, which typically increases as the encoder progresses to capture higher-level abstractions. The kernel slides across the input data, computing a dot product between the kernel weights and the values in each local receptive field. This operation is performed independently for each input channel and then summed across all channels to produce the output feature map. The kernel size K is crucial for defining the receptive field of the model, and its dimensionality depends on the type of input data:

- For 1D data, K ranges from 1 (capturing minimal local patterns) up to the full sequence

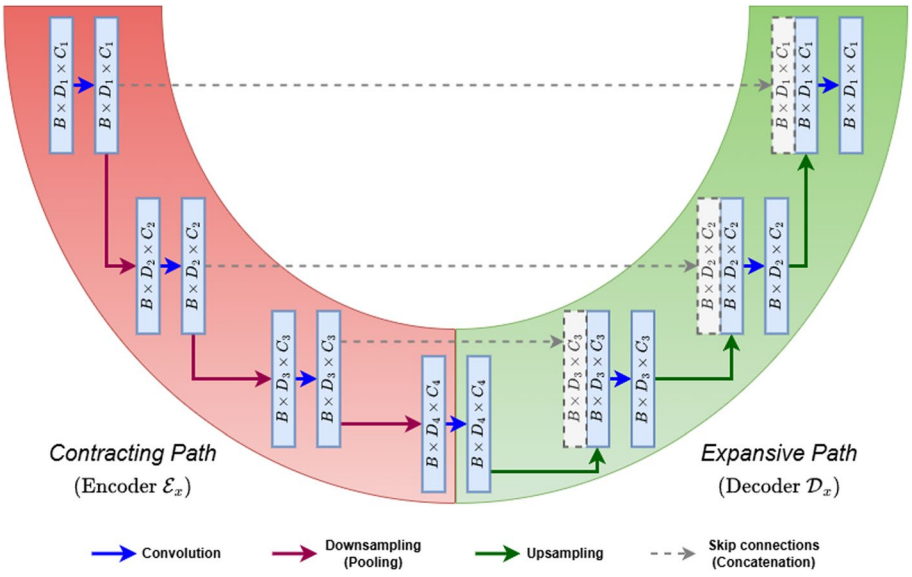

Fig. 2 Diagram of the U-Net architecture. U-Net adopts an encoder–decoder structure with skip connections that link corresponding layers, enabling the preservation of fine-grained details during reconstruction. The contracting path captures hierarchical feature representations, while the expansive path reconstructs the latent representation into the output, supporting high-resolution generative tasks

length L or T.

- For 2D data, such as images, the kernel size is expressed as $K_h \times K_w$, where K_h and K_w denote the height and width of the kernel, respectively.
- For 3D data, such as videos or volumetric inputs, the kernel extends to $K_h \times K_w \times K_t$ for spatiotemporal data (where K_t represents the temporal dimension) or $K_h \times K_w \times K_l$ for volumetric data (where K_l represents depth).

Following the convolution operation, the resulting feature maps are passed through a normalization layer to stabilize and accelerate training. Normalization techniques, such as Batch Normalization (Ioffe and Szegedy 2015), Layer Normalization (Ba et al. 2016), or Instance Normalization (Ulyanov et al. 2016), adjust the distribution of activations to improve gradient propagation and convergence. After normalization, a non-linear activation function is applied element-wise to introduce non-linearity into the model, enabling it to learn complex patterns. The most commonly used activation function in U-Net is the Rectified Linear Unit (ReLU) (Nair and Hinton 2010), defined as:

$$\mathrm{ReLU}(x) = \max(0, x) \tag{1}$$

which retains positive values while setting negative activations to zero.

The final step in each encoder block is pooling, which reduces the spatial or temporal dimensions of the feature map while retaining the most significant information. Pooling achieves this by summarizing local regions, making the model invariant to small translations in the input. The two most common pooling strategies are max pooling, which selects the maximum value within each local region, and average pooling, which computes the mean value. In a max pooling operation with a kernel size $P \times P$ and stride S, the output value $y_{i,j}$ at position (i, j) is computed as the maximum value from the receptive field $R_{i,j}$, which is the region in the input feature map covered by the pooling kernel. The receptive field $R_{i,j}$ for the output position (i, j) corresponds to a $P \times P$ block in the input feature map. Specifically, for a given output position (i, j), the receptive field includes the input values from the region defined by the indices $m \in \{(i-1)S+1, (i-1)S+2, \ldots, (i-1)S+P\}$ and $n \in \{(j-1)S+1, (j-1)S+2, \ldots, (j-1)S+P\}$. The output value $y_{i,j}$ is then determined by the maximum value of $x_{m,n}$ within this receptive field:

$$y_{i,j} = \max_{(m,n)\in R_{i,j}} x_{m,n} \tag{2}$$

where $x_{m,n}$ represents the values in the input feature map, and the operator $\max$ selects the maximum value from the receptive field $R_{i,j}$ of the input. In a 2D setting, a 2×2 max-pooling operation reduces both height and width by half while preserving the most prominent feature in each region. This downsampling process is essential for progressively increasing the receptive field, allowing the encoder to capture global context as the network deepens. Thus, the encoder path effectively compresses the input data into a lower-dimensional representation while preserving essential information through hierarchical feature extraction. The extracted feature maps are then passed to the decoder, where upsampling and refinement occur to reconstruct the output.

2.2.2 Decoder

The second half of the U-Net architecture is the *expansive path*, or decoder, denoted as $\mathcal{D}_x$. Its primary objective is to reconstruct the output while gradually recovering spatial or temporal information lost during the encoding process. Unlike the encoder, which compresses feature representations to extract hierarchical patterns, the decoder progressively increases the resolution of feature maps, ensuring that the final output closely aligns with the original input dimensions. This is accomplished through a combination of upsampling, transposed convolutions, and feature fusion via skip connections.

Each decoder layer consists of two key operations: upsampling and feature refinement. The upsampling process is typically performed using transposed convolution, also known as *deconvolution*, which learns a set of kernels to reverse the effects of downsampling. Given an input feature map $F \in \mathbb{R}^{D' \times C_j}$, where D' represents the reduced spatial or temporal dimensions and C_j is the number of channels, transposed convolution applies a learnable kernel $W \in \mathbb{R}^{K \times C_j \times C_{j-1}}$ to expand the feature map to a higher resolution. This operation increases spatial granularity while maintaining learnable parameters that adapt to the reconstruction process. The transposed convolution can be formally expressed as:

$$F' = W^T * F \tag{3}$$

where $*$ denotes the convolution operation. Unlike traditional interpolation methods (e.g., nearest-neighbor or bilinear upsampling), transposed convolution allows the network to learn optimal upsampling strategies, thereby improving reconstruction fidelity.

To enhance the restoration process, the upsampled feature maps are concatenated with the corresponding feature maps from the encoder through *skip connections*. These connections integrate fine-grained details extracted during encoding with the high-level abstract representations learned by the decoder. This fusion allows the network to recover spatial details that would otherwise be lost due to the progressive downsampling in the encoder.

Each decoder layer applies a standard convolutional block consisting of a convolution operation followed by activation and normalization. The convolution helps refine the upsampled features by learning residual patterns, while the activation function introduces non-linearity to improve expressiveness. The most commonly used activation function at this stage is ReLU. However, alternative activations such as Leaky ReLU (Xu et al. 2015) or Swish (Ramachandran et al. 2018) may be employed depending on the specific application. Batch Normalization or Instance Normalization is often used to stabilize training and mitigate issues such as vanishing gradients.

A crucial aspect of the decoder is its progressive nature. Each layer builds upon the previous one, refining details until the feature maps match the resolution of the original input. The final layer of the decoder typically applies a 1×1 convolution to reduce the number of channels to the desired output dimensionality. For segmentation tasks, this final layer maps the features into a probability distribution over class labels using a softmax or sigmoid activation function. For other applications, such as generative modeling, the final activation may be a hyperbolic tangent (tanh) (Armato et al. 2011) or identity function, depending on the output constraints.

2.2.3 Skip connections

Skip connections are a fundamental component of the U-Net architecture, bridging corresponding encoder and decoder layers to mitigate the loss of spatial information during downsampling. Their primary role is to preserve fine-grained details by directly transferring low-level features from the encoder to the decoder, ensuring that hierarchical information is effectively retained throughout the network. Unlike traditional autoencoder-style architectures that rely solely on progressive downsampling and upsampling, skip connections maintain feature integrity by allowing direct feature reuse, improving both reconstruction accuracy and training stability.

For each level l in the decoder, the upsampled feature map is concatenated with the feature map from the encoder at the same resolution before undergoing further refinement through convolutional operations. This process can be mathematically expressed as:

$$\mathcal{F}_x^{(l)} = \mathrm{Concat}(\mathcal{E}_x^{(l)}, \mathcal{D}_x^{(l)}) \tag{4}$$

where $\mathcal{E}_x^{(l)}$ and $\mathcal{D}_x^{(l)}$ represent the feature maps from the encoder and decoder at layer l, respectively, and $\mathcal{F}_x^{(l)}$ denotes the fused feature map. The concatenated tensor is subsequently processed through convolutional layers to blend high-resolution structural details from the encoder with the abstract semantic representations learned by the decoder.

Beyond simply preserving spatial details, skip connections also facilitate efficient gradient propagation during backpropagation. By establishing direct pathways between early and later layers, they alleviate vanishing gradient issues and accelerate convergence. This is particularly beneficial for training deep networks, where excessive downsampling can otherwise lead to significant information loss and poor reconstruction fidelity. Additionally, they enable the model to leverage multi-scale information, allowing it to capture both local and global features, which is crucial for tasks such as image segmentation and generative modeling.

Require: Input tensor $\mathbf{X}$ (e.g., image), number of encoder levels L
Ensure: Output tensor $\hat{\mathbf{Y}}$ (e.g., segmentation map or generated sample)
Initialize feature map $\mathbf{F}_0 \leftarrow \mathbf{X}$
▷ Encoder: Downsampling path
for $i = 1$ to L **do**
  $\mathbf{F}_i \leftarrow \text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(\mathbf{F}_{i-1})))$
  $\mathbf{F}_i \leftarrow \text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(\mathbf{F}_i)))$
  Store $\mathbf{F}_i$ for skip connection
  if $i < L$ **then**
    $\mathbf{F}_i \leftarrow \text{MaxPool}_{2\times2}(\mathbf{F}_i)$
  end if
end for
▷ Bottleneck
$\mathbf{B} \leftarrow \text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(\mathbf{F}_L)))$
$\mathbf{B} \leftarrow \text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(\mathbf{B})))$
▷ Decoder: Upsampling path
for $j = L$ down to 1 **do**
  $\mathbf{U} \leftarrow \text{UpConv}_{2\times2}(\mathbf{B})$
  $\mathbf{U} \leftarrow \text{Concat}(\mathbf{U}, \mathbf{F}_j)$ ▷ Skip connection
  $\mathbf{U} \leftarrow \text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(\mathbf{U})))$
  $\mathbf{U} \leftarrow \text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(\mathbf{U})))$
  $\mathbf{B} \leftarrow \mathbf{U}$
end for
▷ Final output layer
$\hat{\mathbf{Y}} \leftarrow \phi(\text{Conv}_{1\times1}(\mathbf{B}))$
return $\hat{\mathbf{Y}}$

Algorithm 1 Default U-Net Architecture for Generative Modeling

To guide reproducibility and provide a clear overview of the default U-Net design, Algorithm 1 outlines the canonical forward pass structure. It consists of a symmetric encoder–decoder pathway with skip connections that bridge corresponding feature maps, allowing for the retention of fine-grained spatial details during reconstruction. This standard form underlies most U-Net-based generative architectures before additional modules are integrated.

In summary, the U-Net architecture integrates an encoder–decoder framework with skip connections, making it highly effective for tasks requiring precise spatial reconstruction. The encoder compresses input data into hierarchical feature representations, learning abstract patterns through convolution and downsampling. The decoder reverses this process, progressively reconstructing the input while integrating fine-grained details via upsampling and skip connections. Skip connections play a crucial role in preserving spatial details and improving training efficiency, ensuring that high-resolution outputs can be generated without sacrificing deep semantic understanding. By combining these elements, U-Net achieves a balance between spatial precision and high-level feature extraction, making it a widely adopted model across various domains, including medical imaging, natural image processing, and generative AI.

2.3 Survey methodology

To provide a comprehensive and up-to-date overview of the role of U-Net in generative modeling, this survey follows a structured literature review approach. The objective is to consolidate research efforts that incorporate U-Net into modern generative architectures, spanning multiple data modalities and model types.

Survey landscape As an initial step, we reviewed existing surveys on both Generative AI and U-Net to position our contribution within the broader literature. Table 1 summarizes selected representative survey papers, categorized by their primary focus. Most works

Table 1 Summary of selected survey papers related to Generative AI applications and U-Net-based approaches

Title	Survey focus	Description
A Survey of Generative AI Applications (Gozalo-Brizuela and Garrido Merchan 2024)	Applications	A structured survey of over 350 generative AI applications across multiple domains, highlighting trends and fostering understanding of current developments
Empowering the Metaverse with Generative AI: Survey and Future Directions (Qin and Hui 2023)	Metaverse	Surveys how generative AI can empower metaverse applications, identifies research gaps, and proposes a roadmap for future development
A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT (Cao et al. 2023)	Applications	Reviews the history, techniques, and recent advances in AI-Generated Content (AIGC), covering both unimodal and multimodal models, and discusses future challenges
A Survey on Generative AI and LLM for Video Generation, Understanding, and Streaming (Zhou et al. 2024)	Video	Surveys the impact of Generative AI and large language models (LLMs) on video generation, understanding, and streaming. It highlights achievements, challenges, and future directions in multimedia and AI
Generative AI in Vision: A Survey on Models, Metrics and Applications (Raut and Singh 2024)	Vision	Provides a comprehensive overview of generative AI diffusion models, covering techniques, applications, challenges, and future research directions across text, image, and audio generation
A State-of-the-art Survey of U-Net in Microscopic Image Analysis: from Simple Usage to Structure Mortification (Wu et al. 2022)	U-Net	Surveys the development and improvements of U-Net models for microscopic image segmentation, highlighting technical trends, application areas, and future research directions
Modality specific U-Net variants for biomedical image segmentation: A survey (Punn and Agarwal 2021)	U-Net	Reviews U-Net and its variants for biomedical image segmentation across different modalities, emphasizing their role in disease diagnosis and computer-aided diagnosis systems, including applications to COVID-19
A Survey of Audio Enhancement Algorithms for Music, Speech, Bioacoustics, Biomedical, Industrial, and Environmental Sounds by Image U-Net (Gul and Khan 2023)	U-Net	Surveys U-Net-based approaches for audio enhancement (AE), emphasizing the use of spectrograms as 2D inputs and reviewing diverse applications across speech, music, environmental, biomedical, and industrial sounds
The Role of U-net in the Evolution of Generative AI: Review, Challenges, and Future (ours)	U-Net	Surveys the evolution of U-Net from biomedical segmentation to a key backbone in generative AI, analyzing its adaptations across image, text, audio, video, 3D, and pose/action generation, and highlighting future directions for lightweight and multimodal integration

Each entry highlights the title, survey focus, and a brief description. Unlike previous surveys, our work specifically explores U-Net as a structural backbone in Generative AI, analyzing its roles, adaptations, and significance across modalities

on Generative AI concentrate on applications or specific domains (e.g., metaverse, video, vision), while U-Net surveys primarily target segmentation tasks in biomedical imaging or audio enhancement. These works rarely address U-Net's architectural evolution or its broader role in generative modeling.

Unlike these prior efforts, our survey specifically focuses on U-Net as a structural backbone within generative AI models, analyzing how it has been adapted, extended, and integrated across diffusion, GAN-based, and autoregressive paradigms. This unique focus aims to bridge the gap between architecture-level analysis and generative modality applications.

Scope of literature The survey focuses on peer-reviewed journal articles, preprints, and high-impact conference papers (e.g., NeurIPS, CVPR, ICCV, ICML, ICLR, ACL, AAAI, MICCAI) published from 2015 to 2024. Priority was given to works that apply U-Net to Generative AI tasks across diverse domains, including image generation, text and language modeling, audio and signal processing, video synthesis, 3D data reconstruction, and human pose/action generation. Note that we include segmentation tasks as a subcategory of generative AI, as they involve generating specific modality outputs (e.g., segmentation masks in image segmentation) prior to the subsequent classification or analysis.

Search strategy Relevant literature was identified using academic databases such as Google Scholar, IEEE Xplore, Semantic Scholar, PubMed, and arXiv. Keywords included combinations of: `"U-Net"`, `"generative models"`, `"diffusion"`, `"GAN"`, `"autoregressive"`, `"multi-modal generation"`, `"image generation"`, `"text generation"`, `"audio generation"`, `"video generation"`, `"3D generation"`, and `"pose generation"`. Boolean expressions such as `"U-Net" AND "generative"` or `"U-Net" AND "diffusion" AND "image generation"` were employed to filter domain-specific results.

Inclusion criteria A paper was included if it (1) incorporated U-Net or a U-Net variant in its generative architecture, (2) applied it to one or more data modalities in a generative context, and (3) provided empirical results, visualizations, or theoretical insights regarding performance, design, or architectural advantages. Our methodology categorized these papers based on their generative model family: *diffusion-based*, *GAN-based*, or *autoregressive*, and their targeted modality.

Review organization The selected works are organized thematically in Sect. 4 by data modality, rather than by algorithmic category alone. This modality-centric structure emphasizes U-Net's adaptability across generative contexts. Complementary sections examine architectural adaptations (Sect. 3), advantages (Sect. 6), challenges (Sect. 7), and future directions (Sect. 8).

Note that only a select number of papers are included in this review, and the inclusion is based on their relevance to the key themes and advancements in generative AI involving U-Net architectures. By following this methodology, the survey aims to systematically map the landscape of U-Net-enabled generative models and provide a foundational reference for future research. Figure 3 presents the representative U-Net-based generative models identified through our structured survey methodology, organized by paradigm, data modality, and publication year.

Type/Modality	Image	Text	Audio	Video	3D	Pose/Action
Standard	ResUNet (2018) Shift-Net (2018) MultiResUNet (2020)	U-Net (2018) U-Net Transformer (2019)	Deep U-Net (2017) Wave-U-Net (2018) Attention Wave-U-Net (2019) VQVC+ (2020)	Motion U-Net (2021) AGUNet (2021)	3D U-Net (2016) S3D U-Net (2018) dResU-Net (2022)	Variational U-Net (2018) Reward Driven U-Net (2020) HourGlass U-Net (2020) UNETR-Pose (2021)
Autoregressive	PixelCNN++ (2017)	UET (2024)				
Generative Adversarial Networks	Pix2Pix (2017) PatchGAN (2017) SC-Unet (2024)	U-Net GAN (2024)	UNetGAN (2020)	AP-GAN (2022)	3DGAUnet (2023)	PG² (2017) Attention ResCUNet-GAN (2020)
Diffusion	Stable Diffusion (2022) Imagen (2022) DreamBooth (2022) SmartBrush (2022) ControlNet (2023)		AudioLDM (2023) Stable Audio (2024) Tango 2 (2024)	MagicVideo (2022) Latent-Shift (2023) Lumiere (2024) Upscale-A-Video (2024)	DreamFusion (2022) Efficient-3DiM (2023) Video3D (2024)	DreamPose (2023)

Fig. 3 Representative U-Net-based generative models organized by paradigm, data modality, and publication year. The timeline highlights the progressive evolution of U-Net architectures, ranging from standard encoder–decoder designs to autoregressive, GAN-based, and diffusion-driven approaches, and illustrates their diverse applications across image, text, audio, video, 3D, and pose or action domains

3 Architectural enhancements and model integration with U-Net

While the original U-Net architecture has been widely applied in segmentation tasks, recent adaptations have significantly improved its effectiveness in generative AI. Key modifications include alternative normalization layers to enhance stability, attention mechanisms to capture long-range dependencies, and residual connections to improve information flow. These advancements collectively enable U-Net to generate high-quality outputs while maintaining training efficiency and expressivity. Furthermore, U-Net has been successfully integrated with various generative models, including Diffusion, GANs, and autoregressive models, enhancing its capabilities and allowing it to excel in diverse generative tasks. Table 2 summarizes the key enhancements made to the U-Net architecture, as well as the various generative models, such as Diffusion, GANs, and autoregressive models, with which U-Net has been successfully integrated.

3.1 Architectural enhancements

3.1.1 Normalization layers

Normalization is crucial in deep learning, as it improves training stability and accelerates convergence by standardizing activations across layers. While Batch Normalization (BatchNorm) (Ioffe and Szegedy 2015) has traditionally been dominant in convolutional networks for reducing internal covariate shift (Çiçek et al. 2016; Ibtehaz and Rahman 2020; Shin et al. 2020; tamasino52 2021; Zhang et al. 2018), recent trends in generative models and diffu-

Table 2 Modality-based categorization of U-Net models and their architectural characteristics, including normalization methods, attention mechanisms, and residual connections

Modality	Model	Type	Norm	Attn	ResConn
Image	ResUNet (Zhang et al. 2018)	Standard	BN	None	✓
	Shift-Net (Yan et al. 2018)	Standard	IN	None	×
	MultiResUNet (Ibtehaz and Rahman 2020)	Standard	BN	None	✓
	PixelCNN++ (Salimans et al. 2017)	Autoreg	–	None	✓
	Pix2Pix/PatchGAN (Isola et al. 2017)	GAN	BN	None	×
	SC-Unet (Hou et al. 2024)	GAN	–	Self	×
	Stable Diffusion (Rombach et al. 2022)	Diffusion	LN	Cross	✓
	Imagen (Saharia et al. 2022)	Diffusion	LN	Cross	✓
	DreamBooth (Ruiz et al. 2022)	Diffusion	LN	Cross	✓
	SmartBrush (Xie et al. 2022)	Diffusion	LN	Cross	✓
	ControlNet (Zhang et al. 2023)	Diffusion	LN	Cross	✓
Text	U-Net (Sun et al. 2018)	Standard	–	Self	×
	U-Net Transformer (Donahue et al. 2019)	Standard	LN	Self	×
	UET (Ignacio et al. 2025)	Autoreg	LN	Self	×
	U-Net GAN (Tang and Chen 2024)	GAN	–	None	×
Audio	Deep-U-Net (Jansson et al. 2017)	Standard	BN	None	×
	Wave-U-Net (Macartney and Weyde 2018)	Standard	–	None	×
	Attn Wave-U-Net (Giri et al. 2019)	Standard	–	Self	×
	VQVC+ (Wu et al. 2020)	Standard	IN	None	×
	UNetGAN (Hao et al. 2020)	GAN	BN	None	×
	AudioLDM (Liu et al. 2023)	Diffusion	LN	Cross	✓
	Stable Audio (Evans et al. 2024)	Diffusion	LN	Cross	✓
	Tango 2 (Majumder et al. 2024)	Diffusion	LN	Cross	✓
Video	Motion U-Net (Rahmon et al. 2021)	Standard	–	None	✓
	AGUNet (Yin et al. 2021)	Standard	–	None	×
	AP-GAN (Zhang et al. 2022)	GAN	IN/LN	None	×
	MagicVideo (Zhou et al. 2022)	Diffusion	LN	Tm/Cross	✓
	Latent-Shift (An et al. 2023)	Diffusion	LN	Cross	✓
	Lumiere (Bar-Tal et al. 2024)	Diffusion	–	Self	✓
	Upscale-A-Video (Zhou et al. 2024)	Diffusion	–	Tm/Cross	✓
3D	3D U-Net (Çiçek et al. 2016)	Standard	BN	None	×
	S3D U-Net (Chen et al. 2018)	Standard	IN	None	✓
	dResU-Net (Raza et al. 2022)	Standard	BN	None	✓
	3DGAUnet (Shi et al. 2023)	GAN	BN	None	×
	DreamFusion (Poole et al. 2022)	Diffusion	LN	Cross	✓
	Efficient-3DiM (Jiang et al. 2023)	Diffusion	LN	Cross	✓
	Video3D (Ha 2024)	Diffusion	LN	Cross	✓
Pose/Action	Variational U-Net (Esser et al. 2018)	Standard	–	None	✓
	Reward Driven U-Net (Shin et al. 2020)	Standard	BN	None	×
	HourGlass U-Net (Bulat et al. 2020)	Standard	BN	None	✓
	UNETR-Pose (tamasino52 2021)	Standard	BN	Self	×
	PG^2 (Ma et al. 2017)	GAN	–	None	✓
	Attn ResCUNet-GAN (Na et al. 2020)	GAN	BN	Self	✓
	DreamPose (Karras et al. 2023)	Diffusion	LN	Cross	✓

The comparison highlights the evolution from early models using BatchNorm without attention to diffusion models incorporating LayerNorm, cross-attention, and residual connections, illustrating the architectural progression across modalities

sion architectures have shifted towards alternative techniques, such as Group Normalization (GroupNorm) (Wu and He 2018), Layer Normalization (LayerNorm) (Ba et al. 2016), and Instance Normalization (InstanceNorm) (Ulyanov et al. 2016). These methods offer greater flexibility and stability, particularly in handling small or non-standard batch sizes, and have become increasingly common in state-of-the-art models.

GroupNorm partitions feature channels into smaller groups and computes normalization statistics within each group, eliminating batch size dependence. This makes GroupNorm particularly suitable for tasks where the batch size varies or is small, as it ensures stable training without relying on the entire batch. Similarly, LayerNorm normalizes activations across the feature dimensions of each individual sample, making it a common choice in Diffusion and Transformer-based architectures (Evans et al. 2024; Ha 2024; Jiang et al. 2023; Karras et al. 2023; Majumder et al. 2024; Rombach et al. 2022; Saharia et al. 2022), where maintaining consistent activation distributions across layers is crucial. InstanceNorm, which normalizes each individual sample in a mini-batch, has been particularly successful in style transfer tasks (Wu et al. 2020; Zhang et al. 2022).

BatchNorm computes the mean and variance statistics of activations across the entire mini-batch, which helps stabilize training in conventional networks. The theoretical limitation of BatchNorm lies in its reliance on batch-wide statistics. When training with small or variable batch sizes, the computed statistics may be noisy, leading to unstable training. In contrast, GroupNorm, LayerNorm, and InstanceNorm are less sensitive to batch size and are more suitable for tasks that require flexibility in batch processing. Although empirical studies specifically comparing BatchNorm with these alternatives in generative models may not always be present, the theoretical advantages of these techniques suggest they may mitigate the issues associated with BatchNorm.

3.1.2 Attention mechanisms

Conventional convolutional networks are limited by their reliance on local receptive fields, which hinder their ability to capture long-range dependencies. Integrating attention mechanisms into U-Net has significantly expanded its expressivity in generative AI, where global coherence and fine-grained details are crucial. By incorporating self-attention and cross-attention, U-Net overcomes the limitations of traditional convolutional approaches, enabling it to model complex relationships across images, videos, and other generative tasks with greater flexibility and precision. Figure 4 illustrates architectural designs that combine Transformer and U-Net.

Self-attention, introduced in Transformer architectures (Vaswani et al. 2017), allows the network to dynamically weigh the importance of features across the entire input space, making it particularly effective in high-resolution generative tasks. Models such as U-Net Transformer (Donahue et al. 2019), Attention Wave-U-Net (Giri et al. 2019), and Attention ResCUNet-GAN (Na et al. 2020) integrate self-attention into their U-Net architectures to refine feature representations and capture complex dependencies across dimensions. Self-attention is particularly prominent in unimodal applications, such as Text Generation and Speech Denoising. Most models integrate self-attention inside the convolutional blocks.

Cross-attention mechanisms are essential in conditional generative tasks, such as text-to-image and text-to-audio generation, where they facilitate the alignment of features across different modalities. In these models, cross-attention allows the network to dynamically

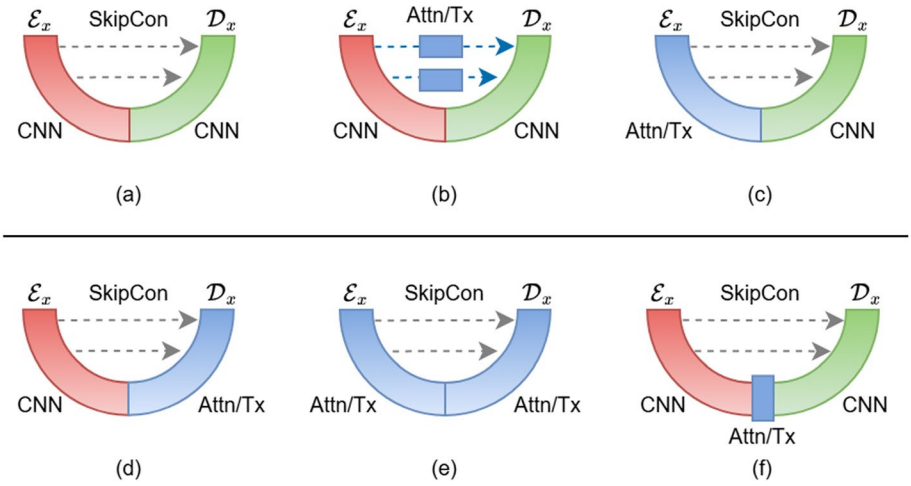

Fig. 4 Integration of Transformer modules into U-Net architectures. **a** Standard U-Net. **b** Skip connections replaced with self-attention or Transformer modules. **c** Transformer modules integrated into the encoder. **d** Transformer modules integrated into the decoder. **e** Transformer modules used in both encoder and decoder, forming a fully Transformer-based U-Net. **f** Transformer modules inserted between encoder and decoder, as in the U-Net Encapsulated Transformer (UET)

attend to relevant parts of the input from one modality (e.g., text) and integrate it into the feature space of another modality (e.g., image or audio). This enables the model to condition image or audio generation on external inputs, such as text embeddings, ensuring that the generated outputs align with the provided context. For example, in text-to-image tasks (Rombach et al. 2022; Ruiz et al. 2022; Saharia et al. 2022; Zhang et al. 2023), cross-attention helps generate high-quality images that are consistent with given text descriptions, while in text-to-audio tasks (Evans et al. 2024; Liu et al. 2023; Majumder et al. 2024), it ensures that audio features correspond to textual inputs. The advantages of cross-attention include improved coherence between input and output, better handling of complex dependencies, and enhanced interpretability, as it provides a clear mapping between the conditioning context and the generated content.

An emerging adaptation of attention mechanisms within U-Net architectures involves encapsulating Transformer modules, wherein the Transformer encoder is employed to capture the full contextual representation of the data (Ignacio et al. 2024). At the same time, the Transformer decoder is utilized for autoregressive tasks (Wang et al. 2024). This U-Net Encapsulated Transformer (UET) architecture introduces an optimization strategy that reduces the embedding dimension before applying self-attention, thereby mitigating the computational complexity associated with standard Transformer models. Studies have shown that UET effectively balances computational efficiency and model performance, demonstrating its potential in tasks such as classification and recommendation while maintaining competitive accuracy despite dimensionality reduction. A promising application of the UET architecture in language modeling demonstrates its potential to improve both efficiency and performance in generative tasks (Ignacio et al. 2025).

Other related works in natural language processing incorporating self-attention in U-Net are used in reading comprehension (Sun et al. 2018) and dialogue generation tasks (Donahue

et al. 2019). Moreover, attention-enhanced U-Net variants have proven effective in speech and audio processing tasks. For instance, Attention Wave-U-Net (Giri et al. 2019) integrates a local self-attention mechanism within the Wave-U-Net architecture for speech enhancement. By applying the attention mechanism to the skip connections, it processes raw waveforms directly and is trained end-to-end. This inclusion significantly improves performance in terms of speech quality metrics, outperforming other speech enhancement approaches on the Voice Bank Corpus (VCTK) dataset. The Attention ResCUNet-GAN model integrates U-Net with attention gates to enhance pose-invariant face recognition. This model uses coupled U-Nets with attention gates in the skip connections, refining feature propagation and ensuring robustness to variations in facial poses. By learning to identify faces across different poses, it generates accurate face representations regardless of the pose, improving recognition performance. Experimental results on benchmarks such as Multi-PIE and LFW demonstrate superior performance compared to existing methods (Na et al. 2020).

3.1.3 Residual connections

Residual connections, first introduced in Residual Networks (ResNet) (He et al. 2016), are widely used in U-Net-based generative models to address the challenge of vanishing gradients in deep networks. By introducing shortcut pathways that bypass intermediate layers, residual connections improve gradient flow, ensuring effective feature preservation and enhancing training stability. This is particularly valuable in generative tasks across modalities, where maintaining high-fidelity details throughout the iterative generation process is crucial.

Residual blocks, when integrated into convolutional layers in U-Net architectures, facilitate the transformation of feature maps while preserving essential spatial, temporal, or structural information. This design enhances gradient flow, mitigates degradation in deep networks, and improves feature reuse, allowing models to generate high-quality outputs with fine-grained details. The incorporation of residual blocks has been widely adopted across various generative tasks, from image segmentation (Zhang et al. 2018) to video generation (An et al. 2023; Bar-Tal et al. 2024; Zhou et al. 2022, 2024) and 3D reconstruction (Chen et al. 2018; Raza et al. 2022), where it supports stable training and improves the fidelity of generated features. MultiResUNet (Ibtehaz and Rahman 2020) introduces ResNet-like skip connections within the U-Net, incorporating residual pathways called Res Paths to refine feature propagation and maintain hierarchical representations. This approach enhances the model's ability to learn complex patterns, improving tasks like medical image segmentation where both fine details and global context are crucial.

The use of residual connections has become a defining characteristic of modern diffusion models (Evans et al. 2024; Jiang et al. 2023; Liu et al. 2023; Majumder et al. 2024; Poole et al. 2022; Rombach et al. 2022; Ruiz et al. 2022; Saharia et al. 2022; Zhang et al. 2023), regardless of their modalities. In practice, generative U-Net architectures often stack multiple residual blocks alongside attention mechanisms, such as self-attention or temporal attention, to maximize both stability and the fidelity of generated outputs. These enhancements enable residual-enhanced U-Net variants to effectively handle high-resolution, high-fidelity synthesis across various modalities, including spatial, temporal, and structural data, making them well-suited for complex generative modeling tasks.

3.2 Model integration with U-Net

3.2.1 Diffusion models

Diffusion models have revolutionized generative AI by introducing an iterative denoising process that progressively transforms pure noise into high-quality outputs. Unlike GANs, which rely on adversarial training, diffusion models operate by sequentially perturbing data with noise and then learning to reverse the process. At the core of these models, U-Net serves as the denoising function, predicting the noise component at each step and enabling high-fidelity reconstruction. The versatility of U-Net within diffusion frameworks extends beyond image generation to applications in audio, video and 3D modalities, demonstrating its adaptability across various generative tasks.

U-Net architectures have become central to diffusion models, particularly in the context of high-quality image synthesis. Models such as Stable Diffusion (Rombach et al. 2022), Imagen (Saharia et al. 2022), ControlNet (Zhang et al. 2023), SmartBrush (Xie et al. 2022), and DreamBooth (Ruiz et al. 2022) utilize the U-Net module, which refines images through iterative denoising. Stable Diffusion employs a latent-space U-Net to process compressed image representations, thereby enhancing computational efficiency while preserving image details. Imagen incorporates cross-attention layers within the U-Net, enabling precise text-to-image synthesis. These innovations enhance diffusion models over GANs, addressing issues such as mode collapse and training instability while producing high-resolution, diverse, and photorealistic outputs. The capabilities of U-Net have also been extended to 3D generation. Models such as Efficient-3DiM (Jiang et al. 2023), DreamFusion (Poole et al. 2022), and Video3D (Ha 2024) utilize U-Net as a denoising backbone to generate novel views of 3D shapes from a single image. Through iterative refinement of noisy data, the framework produces high-quality, view-consistent 3D representations.

Diffusion models have also made significant strides in video generation, where maintaining both spatial and temporal consistency is crucial for effective results. U-Net has been adapted with 3D convolutions and spatiotemporal attention mechanisms to model motion dynamics across frames. For example, the Upscale-A-Video framework (Zhou et al. 2024) incorporates 3D convolutions and temporal attention to reduce flickering and ensure stability. Models such as the Lumiere (Bar-Tal et al. 2024) generate video durations in a single pass, thereby improving global temporal consistency. The Latent-Shift model (An et al. 2023) introduces a parameter-free temporal shift module to extend image generation to video generation, efficiently learning motion across temporal dimensions. Similarly, Video3D (Ha 2024) uses a video diffusion model and neural volume renderer to convert a single image into a 3D scene, generating temporally coherent multi-view representations. Finally, the MagicVideo framework (Zhou et al. 2022) employs a 3D U-Net within a latent diffusion model for efficient text-to-video generation, ensuring smooth and coherent outputs. These innovations have established diffusion-based video generation models as key tools in AI-generated animations, video upscaling, and frame interpolation, where U-Net contributes to both high-resolution rendering and motion continuity.

3.2.2 Generative adversarial networks

Generative Adversarial Networks (GANs) comprise a generator and a discriminator operating within a min-max optimization framework. The generator synthesizes data, while the discriminator evaluates its authenticity, refining its ability to distinguish between real and generated instances. U-Net has been integrated into both components to enhance image synthesis and feature propagation, owing to its hierarchical structure and skip connections, which preserve fine-grained details.

U-Net as a Generator In GANs, U-Net serves as the generator, providing a structured approach to refining output synthesis. Unlike traditional GANs, which use transposed convolutions for upscaling, U-Net's encoder–decoder structure with skip connections ensures that spatial features are directly propagated, thereby maintaining the input structure throughout the synthesis. This design has proven effective across various generative tasks, including image generation (Isola et al. 2017) and even speech enhancement (Hao et al. 2020). In applications such as pose-guided synthesis (Ma et al. 2017), U-Net has demonstrated its ability to generate coherent outputs by retaining detailed spatial features and enhancing structural integrity.

U-Net as a Discriminator As a discriminator, U-Net operates on a pixel-wise level, offering more granular feedback compared to traditional discriminators that assess images holistically. For example, PatchGAN (Isola et al. 2017) utilizes the U-Net architecture to enhance the discriminator's ability to capture fine details. Additionally, the incorporation of U-Net into the discriminator enhances GAN stability by improving gradient flow through its skip connections. This prevents issues like mode collapse, promoting more reliable and diverse output generation. These advantages make U-Net a valuable component in both the generator and discriminator of GANs, enhancing model performance across multiple modalities, including images, audio, and 3D data.

U-Net as a Generator and Discriminator in GANs In GANs, U-Net has been successfully integrated into both the generator and discriminator roles to enhance image synthesis and inpainting tasks. One such example is a U-Net-based architecture that incorporates the Wasserstein Generative Adversarial Network (WGAN) for image inpainting (Hou et al. 2024). In this model, the generator is a Symmetric Connected U-Net (SC-Unet), which uses skip connections between every encoder and decoder block to maintain a balanced, symmetrical architecture. By utilizing dilated convolutions and multi-head self-attention (MHSA) in selected blocks, the model enhances feature propagation and captures long-range spatial dependencies, enabling seamless inpainting of missing regions. The generator benefits from the more granular feedback provided by the discriminator, which, unlike traditional GANs, operates at the pixel level. The discriminator calculates the probability of each pixel belonging to either the hole or non-hole region of the image, providing the generator with richer gradient loss information for more detailed inpainting. This architecture demonstrates significant improvements in image quality, as shown in evaluations on datasets like CelebA-HQ and ImageNet, with notable improvements in metrics like PSNR and SSIM. The combination of U-Net as both a generator and a discriminator in this context enables more accurate and visually plausible results, particularly for image inpainting tasks.

3.2.3 Autoregressive models

Autoregressive models generate data sequentially, predicting each element based on previously generated outputs. While transformer-based architectures like GPT (Radford et al. 2018) and PixelCNN (Oord et al. 2016) have been dominant in tasks like text generation and image synthesis, U-Net has emerged as a complementary module to enhance spatial coherence and local feature modeling in autoregressive tasks. An example is PixelCNN++ (Salimans et al. 2017), which incorporates downsampling and shortcut connections similar to U-Net to improve image generation and accelerate training. A promising application of U-Net as an autoregressive model is the U-Net Encapsulated Transformer (UET) (Ignacio et al. 2024; Wang et al. 2024), which serves as an effective next-token predictor. This framework has shown potential in language modeling tasks (Ignacio et al. 2025), where UET reduces the computational complexity of the Transformer model.

A primary challenge in applying autoregressive models across different modalities is the sequential nature of their output generation, which contrasts with the whole-output generation of U-Net. This discrepancy makes it difficult to directly incorporate U-Net into autoregressive tasks. Moreover, autoregressive models often struggle to capture long-range dependencies and global context, a limitation inherent in traditional approaches, such as RNNs. However, this challenge has been partially addressed by Transformer models, which excel at capturing such dependencies. UET exemplifies a hybrid solution, combining U-Net's hierarchical feature extraction capabilities with the global context modeling of Transformers, enabling autoregressive training and inference. This integration enables the model to preserve fine-grained details while effectively learning both local and long-range dependencies, thereby facilitating improved performance in tasks such as next token prediction and language modeling.

4 Modality-centric applications of U-Net in generative AI

The widespread success of U-Net in generative modeling stems from its flexibility across a broad range of data modalities. This section highlights the diverse applications of U-Net in generative tasks across multiple data types. It encompasses a range of applications, including image synthesis and text-to-image generation, as well as audio signal processing, video synthesis, 3D volumetric reconstruction, and the generation of human poses or actions. By examining these modality-specific use cases, we emphasize U-Net's continued relevance in shaping state-of-the-art generative systems and its growing role in the multimodal landscape of modern AI. Table 3 summarizes the list of models with their corresponding applications.

4.1 Image generation

Among the various modalities in Generative AI, image generation remains the most mature and widely explored domain, where U-Net has demonstrated exceptional performance. As a foundational architecture in modern generative pipelines, U-Net plays a central role in modeling both local textures and global structures. This is crucial for photorealistic image synthesis and controllable image editing. Its encoder–decoder framework with skip connections makes it particularly suitable for iterative refinement processes, which are central

Table 3 Summary of U-Net applications in generative AI

Category	Application	Model	Description
Vision	Image Generation	PixelCNN++ (Salimans et al. 2017)	Used a discretized logistic mixture likelihood, downsampling for multi-resolution feature extraction, and incorporated shortcut connections for faster optimization
		Pix2Pix/ PatchGAN (Isola et al. 2017)	A conditional adversarial network for image-to-image translation, where the model learns a mapping from input to output images along with an adaptive loss function
	Text-to-Image	Stable Diffusion (Rombach et al. 2022)	Leverages latent diffusion models (LDMs) to efficiently train and generate high-quality images in the latent space of pretrained autoencoders, significantly reducing computational costs
		Imagen (Saharia et al. 2022)	A text-to-image diffusion model that achieves state-of-the-art photorealism and text-image alignment by leveraging large pretrained language models for text encoding
		DreamBooth (Ruiz et al. 2022)	Enables tasks like subject recontextualization and text-guided view synthesis with minimal input data by leveraging a unique identifier for the subject and a class-specific prior preservation loss
		ControlNet (Zhang et al. 2023)	Extends pretrained text-to-image diffusion models by incorporating spatial conditioning controls, such as edges, depth, and human pose, using "zero convolutions" to prevent harmful noise during fine-tuning
	Image Inpainting	Shift-Net (Yan et al. 2018)	Improves the accuracy and realism of inpainted images by shifting the encoder features from the known region and applying a guidance loss
		SC-Unet (Hou et al. 2024)	Integrates the Wasserstein Generative Adversarial Network (WGAN) with a symmetric U-Net architecture using skip connections to enhance image reconstruction
		SmartBrush (Xie et al. 2022)	Incorporates object-mask prediction and multi-task training with text-to-image generation, maintaining better background preservation and allowing shape control for inpainted objects
	X-to-Video	MagicVideo (Zhou et al. 2022)	Utilizes a pre-trained VAE to map video clips into a low-dimensional latent space, where a diffusion model learns the video distribution and incorporates novel techniques such as a frame-wise lightweight adaptor and directed temporal attention to capture temporal dependencies
		Lumiere (Bar-Tal et al. 2024)	Introduces Space-Time U-Net, which generates an entire video in a single pass, addressing challenges associated with temporal consistency in traditional methods that generate keyframes followed by temporal super-resolution

Table 3 (continued)

Category	Application	Model	Description
Vision	Text-to-Video	Latent-Shift (An et al. 2023)	Models temporal information by shifting feature map channels along the temporal dimension without adding new parameters, making it more computationally efficient than other video generation methods
	Video Synthesis	AP-GAN (Zhang et al. 2022)	Introduces identity (ID) and pose-expression (PE) blocks in the generator, along with a PE-aware discriminator and perceptual loss to ensure attribute consistency and realism across video frames
	Video Upscaling	Upscale-A-Video (Zhou et al. 2024)	Enhances temporal consistency using local temporal layers in the U-Net and decoder, as well as a global flow-guided latent propagation module
	Text-to-3D	DreamFusion (Poole et al. 2022)	Performs text-to-3D generation by leveraging a pre-trained 2D diffusion model and optimizing a Neural Radiance Field (NeRF) to match the 2D output via score distillation
	2D-to-3D	Efficient-3DiM (Jiang et al. 2023)	Introduces a timestep sampling strategy, a Vision Transformer-based 3D feature extractor, and an improved training scheme to achieve high-quality multi-view outputs with significantly reduced computational overhead
		Video3D (Ha 2024)	Combines a pre-trained video diffusion model and a view-conditioned U-Net with an added voxel grid and volume rendering layer to generate smooth, view-consistent 3D outputs
Text	Machine Translation	U-Net GAN (Tang and Chen 2024)	Integrates a U-Net architecture into a generative adversarial network framework to improve machine translation of long English sentences, enhancing both accuracy and fluency through adversarial training
	Reading Comprehension	U-Net (Sun et al. 2018)	An end-to-end model for machine reading comprehension with unanswerable questions, using a universal node to fuse question and passage information
	Language Modeling	U-Net Transformer (Donahue et al. 2019)	Introduces hierarchical processing into the Transformer architecture by incorporating U-Net-inspired local and skip connections
		UET (Ignacio et al. 2025)	Applies 1D convolution-based dimensionality reduction to token embeddings, enabling deeper and more efficient Transformer-based language models
Audio	Text-To-Audio	AudioLDM (Liu et al. 2023)	Leverages contrastive language-audio pretraining (CLAP) and latent diffusion models to generate high-quality audio from text prompts efficiently
		Stable Audio (Evans et al. 2024)	A latent diffusion model designed for efficient generation of long-form, variable-length 44.1kHz stereo audio from text and timing prompts
		Tango 2 (Majumder et al. 2024)	Improves text-to-audio generation by fine-tuning a diffusion-based model with direct preference optimization (DPO), using a synthetic dataset (Audio-alpaca) of preferred and flawed audio samples

Table 3 (continued)

Category	Application	Model	Description
Audio	Denoising	Deep-U-Net (Jansson et al. 2017)	Applies the U-Net architecture to singing voice separation, using skip connections to preserve fine audio details during source decomposition
		Wave-U-Net (Macartney and Weyde 2018)	A time-domain U-Net variant designed for end-to-end speech enhancement, capable of modeling phase information and leveraging large temporal contexts
		Attn Wave-U-Net (Giri et al. 2019)	Enhances raw waveform speech using a U-Net architecture augmented with an attention mechanism, which improves quality by focusing on salient voice regions
		UNetGAN (Hao et al. 2020)	A time-domain speech enhancement model that combines a U-Net generator with adversarial training, using dilated convolutions to handle extremely low SNR conditions
	Speech Synthesis	VQVC+ (Wu et al. 2020)	A one-shot voice conversion model that combines vector quantization with a U-Net autoencoder architecture to disentangle speaker and content information
Action	Pose Estimation	Variational U-Net (Esser et al. 2018)	Combines a variational autoencoder for modeling appearance with a U-Net architecture for preserving object shape, enabling shape-guided image generation and flexible appearance transfer
		HourGlass U-Net (Bulat et al. 2020)	A hybrid architecture that fuses the HourGlass and U-Net designs, introducing gated per-channel skip connections for efficient data flow control
		UNETR-Pose (tamasino52 2021)	It adapts volumetric segmentation techniques to human pose prediction by leveraging the transformer encoder for global context and U-Net-style skip connections for spatial precision
	Pose Generation	PG (Ma et al. 2017)	A two-stage U-Net-based framework for synthesizing person images in arbitrary target poses. The first stage generates a coarse image using pose integration, while the second stage refines the result with adversarial training to enhance detail and realism
		DreamPose (Karras et al. 2023)	A diffusion-based video generation framework that animates still fashion images using pose sequences, extending Stable Diffusion with pose and image conditioning to produce realistic clothing and human motion
	Face Recognition	Attn ResCUNet-GAN (Na et al. 2020)	A generative model designed for pose-invariant face recognition by completing facial UV maps using a pair of attention-enhanced U-Nets with residual and feature fusion mechanisms
	Navigation	Reward Driven U-Net (Shin et al. 2020)	Integrates a U-Net-based segmentation model with an actor-critic reinforcement learning framework to enable monocular RGB-based obstacle avoidance for drones

This table categorizes various U-Net-based models across different domains, including vision, text, audio, and action generation. Models are organized by their application type and further described by their specific use case and architecture

to diffusion-based generative models. Moreover, recent surveys (Umirzakova et al. 2024) emphasize the importance of upstream image restoration and super-resolution processes in improving the quality of input data, which in turn can significantly impact the performance of downstream generative architectures, such as U-Net.

One of the most transformative innovations in this space is the development of diffusion models, which generate images through a two-phase process: forward diffusion (adding noise) and reverse denoising. Instead of adversarial training, these models progressively corrupt an image and then train a model to reverse this corruption. The U-Net architecture is employed as the denoising function in the reverse process, learning to predict and remove noise at each timestep. In prominent diffusion models like Stable Diffusion (Rombach et al. 2022), Imagen (Saharia et al. 2022), ControlNet (Zhang et al. 2023), and DreamBooth (Ruiz et al. 2022), U-Net serves as the core neural backbone responsible for reconstructing high-fidelity visual outputs from noisy latent representations.

U-Net has also been widely applied in image-to-image translation tasks, such as in Pix-2Pix (Isola et al. 2017), where it serves as the generator in a GAN-based framework. U-Net transforms input images into corresponding target representations, enabling tasks such as sketch-to-photo conversion and domain adaptation. The architecture's skip connections help preserve spatial coherence and fine-grained features during translation, improving output quality and consistency. Similarly, PatchGAN (Isola et al. 2017) utilizes U-Net in the discriminator role for image-to-image translation tasks, where it evaluates image patches rather than whole images, improving fine-grained texture analysis and reducing artifacts. This enables the model to focus on local details, which is particularly useful for applications that require high-quality, localized image editing. Additionally, PixelCNN++ (Salimans et al. 2017), which shares architectural similarities with U-Net, incorporates downsampling and shortcut connections to enhance image generation. These modifications enable the model to capture fine details at multiple resolutions, addressing the challenge of generating high-quality images and improving performance over traditional methods.

U-Net enables advanced image editing capabilities within different frameworks. Through conditioning mechanisms, such as masked inputs, structural guidance, or latent code manipulation, U-Net facilitates localized editing operations without compromising global consistency. Inpainting (Hou et al. 2024; Xie et al. 2022; Yan et al. 2018), for example, fills missing regions using context-aware synthesis, while outpainting coherently extends image boundaries. Other applications include style transfer, where the model is conditioned to change the image's visual style while retaining its content structure. These editing techniques benefit from U-Net's skip connections, which help preserve spatial details during the denoising process.

Based on the representative models summarized in Table 4, the approximate parameter size of a standard U-Net architecture typically remains below 30 million parameters, while GAN-based variants average around 200 million, and diffusion-based models often exceed 400 million parameters. The choice of dataset generally depends on the application domain; medical and remote sensing tasks favor specialized datasets, whereas common image generation studies frequently employ large-scale benchmarks such as ImageNet, CelebA, MS-COCO, and ADE20K. Evaluation metrics also vary across works, with Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Fréchet Inception Distance (FID) being the most widely used for assessing image quality and perceptual realism. However, it is important to note that direct quantitative comparison between models is challenging, as many studies differ in dataset composition, evaluation protocol, and training setup.

Table 4 Quantitative comparison of U-Net-based generative models across vision modalities

Modality	Model	Type	Params	Dataset	FID
Image	ResUNet (Zhang et al. 2018)	Standard	7.8M	Massachusetts roads dataset	–
	MultiResU-Net (Ibtehaz and Rahman 2020)	Standard	18.6M	Medical image datasets	–
	SC-Unet (Hou et al. 2024)	GAN	213 M	ImageNet 2012, CelebA-HQ	8.15, 1.69
	Stable Diffusion (Rombach et al. 2022)	Diffusion	400 M	ImageNet 2012, CelebA-HQ	3.6, 5.11
	Imagen (Saharia et al. 2022)	Diffusion	600 M	MS-COCO	7.27
	SmartBrush (Xie et al. 2022)	Diffusion	–	MS-COCO	5.76
	ControlNet (Zhang et al. 2023)	Diffusion	–	ADE20K	15.27
Modality	Model	Type	Params	Dataset	FVD
Video	Motion U-Net (Rahmon et al. 2021)	Standard	17.8M	CDnet-2014	–
	MagicVideo (Zhou et al. 2022)	Diffusion	–	MSR-VTT, UCF-101	655
	Latent-Shift (An et al. 2023)	Diffusion	–	MSR-VTT, UCF-101	358.34
	Lumiere (Bar-Tal et al. 2024)	Diffusion	–	UCF-101	332.49
Modality	Model	Type	Params	Dataset	DICE
3D	3D U-Net (Çiçek et al. 2016)	Standard	19 M	Xenopus kidney dataset	–
	S3D U-Net (Chen et al. 2018)	Standard	–	BraTS 2018	0.78
	dResU-Net (Raza et al. 2022)	Standard	30.5M	BraTS 2020	0.84

The table presents parameter counts and representative performance metrics, including FID for image generation, FVD for video synthesis, and DICE for 3D reconstruction. The reported values are indicative and not directly comparable across studies, as experimental settings and datasets vary among the referenced works. Missing values indicate results that were not reported in the original papers

4.2 Text and language generation

Text generation, a central task in natural language processing (NLP), has traditionally been dominated by Transformer-based architectures such as GPT (Brown et al. 2020) and BERT (Devlin et al. 2019). These models excel at capturing long-range dependencies and gener-

ating coherent sequences. However, recent efforts have explored how convolutional and U-Net-inspired models can contribute to or complement these architectures, particularly in improving training efficiency, modeling local dependencies, and reducing computational cost in large language models (LLMs) (Ignacio et al. 2025). While GAN-based models such as Adversarial-NMT (Wu et al. 2017) and BR-CSGAN (Yang et al. 2017) have been used for text generation, they primarily rely on RNN-based architectures rather than U-Net structures. On the other hand, state-of-the-art models like ChatGPT (OpenAI 2023), Gemini (Team 2024), and DeepSeek (DeepSeek-AI 2024), which are built on Transformer architectures, have significantly advanced the field of text generation. These models, especially in their autoregressive configurations, continue to push the boundaries of NLP tasks, offering high-quality, human-like text generation across a wide range of applications.

U-Net-inspired models have been applied in NLP for tasks such as machine reading comprehension and dialogue systems. The U-Net model (Sun et al. 2018) tackles machine reading comprehension with unanswerable questions by integrating an answer pointer, a no-answer pointer, and an answer verifier. It processes the question and passage as a unified sequence using a universal node, enabling end-to-end training and improved efficiency. On the SQuAD 2.0 dataset, U-Net achieved an F1 score of 71.7, effectively predicting unanswerable questions. A variant of the Transformer architecture, named U-Net Transformer (Donahue et al. 2019), integrates the hierarchical and local connections from U-Net to produce more accurate abstract representations of input sequences. The proposed model is evaluated on dialogue generation tasks, demonstrating improved performance in terms of perplexity and certain embedding measures, along with increased diversity in the generated responses, compared to the standard Transformer.

One notable adaptation in this direction is the U-Net Encapsulated Transformer (UET) (Ignacio et al. 2024; Wang et al. 2024), a hybrid architecture incorporating U-Net principles into the Transformer framework. Rather than feeding token embeddings directly into a full-resolution transformer, UET first processes the embeddings through a contracting U-Net encoder to reduce dimensionality. This compressed representation is then passed to a lightweight transformer block, which models global context. The resulting features are subsequently upsampled through a U-Net decoder to restore sequence resolution. This strategy significantly reduces the number of trainable parameters and memory usage, making transformer training more feasible under constrained resources. While these models were originally applied to classification and recommendation tasks, their architectural advantages point to broader applicability in language modeling (Ignacio et al. 2025). Figure 5 shows the overview diagram of UET for language modeling.

In the text domain, U-Net and its transformer-augmented variants are primarily applied to language modeling and text generation tasks. These models typically operate on token embeddings rather than pixel or waveform data, with parameter sizes varying widely depending on the inclusion of transformer or attention blocks. While smaller U-Net-based text generators can have tens of millions of parameters, transformer-integrated variants, such as U-Net Encapsulated Transformer (UET), often reach over one hundred million. Datasets for evaluation differ across studies but generally include standard language modeling corpora such as WikiText or large-scale instruction-tuning datasets. The most common evaluation metrics include Perplexity (PPL), BLEU, and ROUGE, depending on whether the model focuses on next-token prediction or controlled text synthesis. Although not directly comparable to image or audio tasks, these benchmarks demonstrate how U-Net's

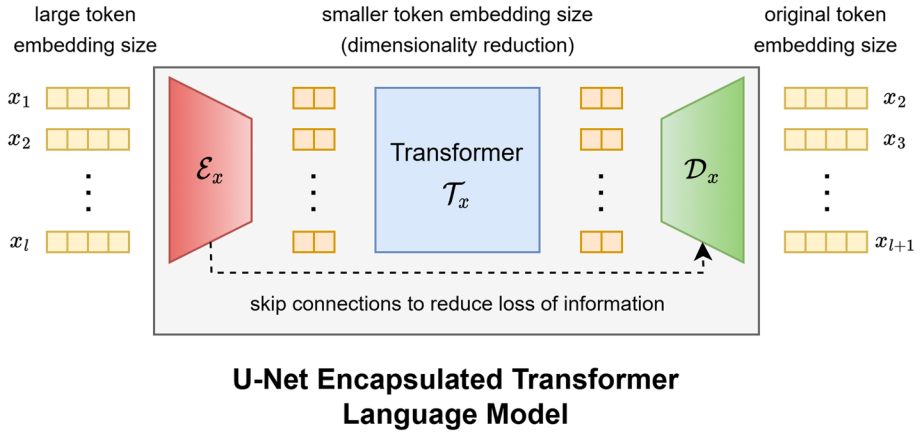

Fig. 5 Illustration of UET for language modeling. High-dimensional token embeddings are compressed by an encoder, processed by a Transformer to learn contextual representations, and then restored by a decoder, with skip connections preserving fine-grained information. This dimensionality reduction during contextual learning lowers computational costs and enables more scalable language modeling

hierarchical encoder–decoder structure supports token-level generation with reduced computational overhead compared to full-scale transformer architectures.

4.3 Audio and speech generation

U-Net architectures have been increasingly adopted in the domain of audio and speech generation due to their ability to capture both local and global features through hierarchical processing. The encoder–decoder structure, combined with skip connections, facilitates the preservation of fine-grained temporal structures and enables the reconstruction of high-fidelity signals from degraded or noisy inputs. These characteristics have made U-Net highly suitable for various generative tasks involving speech and audio signals, such as speech enhancement, source separation, and voice conversion.

In speech enhancement, U-Net models have been shown to effectively recover clean speech from noisy environments. One notable example is the Wave-U-Net architecture (Macartney and Weyde 2018), which operates directly in the time domain, allowing for integrated modeling of phase information and consideration of large temporal contexts. U-Net-based models have also been applied in singing voice separation, where high-resolution temporal detail is critical. One work proposed a deep U-Net architecture for separating vocals from polyphonic music, leveraging the model's ability to retain spatially aligned features through skip connections (Jansson et al. 2017). The architecture proved effective in preserving vocal clarity while minimizing interference from accompaniment, illustrating U-Net's strength in time-frequency processing. Voice conversion is another area where U-Net has shown promise. The VQVC+ model (Wu et al. 2020) employs a vector-quantized U-Net architecture for one-shot voice conversion, where a speaker's vocal characteristics are modified to resemble another's while preserving the linguistic content. This is achieved by disentangling speaker and content information and leveraging U-Net's structured feature propagation to perform robust, high-quality transformations.

Furthermore, U-Net has been integrated into GAN frameworks to enhance speech generation under low signal-to-noise conditions. UNetGAN (Hao et al. 2020) integrates U-Net with dilated convolution operations within a GAN-based framework to enhance speech in extremely low SNR conditions, leveraging the strengths of U-Net for time-domain processing, GANs for improved generator performance, and dilated convolutions to capture large temporal contexts for superior speech enhancement. This approach highlights the potential of combining U-Net architectures with adversarial training to enhance speech quality. Attention-enhanced variants such as the Attention Wave-U-Net (Giri et al. 2019) extend this architecture by incorporating local self-attention mechanisms, improving the model's ability to focus on critical segments of the waveform while suppressing noise. This model processes raw waveforms directly and has achieved state-of-the-art results in speech denoising tasks.

Recent advancements in diffusion models have significantly enhanced text-to-audio (TTA) generation capabilities. AudioLDM (Liu et al. 2023) introduces a latent diffusion framework conditioned on contrastive language-audio pretraining (CLAP) embeddings, enabling high-quality audio synthesis with improved computational efficiency. Building upon this, Fast Timing-Conditioned Latent Audio Diffusion (Stable Audio) (Evans et al. 2024) incorporates timing embeddings, allowing for precise control over audio duration and structure, thus facilitating the generation of long-form, variable-length stereo audio from text prompts. Tango 2 (Majumder et al. 2024) further refines this approach by employing direct preference optimization (DPO) on a synthetic preference dataset, enhancing the alignment between generated audio and textual descriptions, and outperforming previous models in both automatic and manual evaluation metrics. Collectively, these models represent significant strides in the field of generative audio, offering improved fidelity, control, and alignment in TTA tasks.

In the audio and speech domain, U-Net architectures have been widely applied to both speech enhancement and general audio generation tasks. For speech generation and enhancement, the most common datasets are VCTK and SEGAN, which provide clean and noisy speech samples for supervised training. Evaluation typically employs perceptual and signal-based metrics such as Perceptual Evaluation of Speech Quality (PESQ), Composite Signal Distortion (CSIG), Background Noise Intrusiveness (CBAK), Overall Speech Quality (COVL), and Segmental Signal-to-Noise Ratio (SSNR). In contrast, audio generation models use diverse datasets depending on the task, often involving multi-source or curated audio collections. These models are predominantly diffusion-based and exhibit significantly larger parameter sizes, ranging from approximately 739 million to 1 billion parameters. Their performance is commonly assessed using Kullback–Leibler (KL) divergence and CLAP (Contrastive Language-Audio Pretraining) scores to measure distributional similarity and text–audio alignment. Representative parameter sizes and performance metrics for these models are summarized in Table 5.

4.4 Video generation

U-Net architectures have been widely adapted for video generation across various model types, each leveraging U-Net's hierarchical structure and skip connections to maintain both spatial coherence and temporal consistency. In standard U-Net models, attention mechanisms are often integrated to capture long-range dependencies and improve the quality of

Table 5 Quantitative overview of U-Net-based models for audio generation and enhancement

Model	Params	Dataset	PESQ	CSIG	CBAK	COVL	SSNR
Wave-U-Net (Macartney and Weyde 2018)	–	VCTK, SEGAN	2.40	3.52	3.24	2.96	9.97
Attn Wave-U-Net (Giri et al. 2019)	–	VCTK	2.62	3.91	3.35	3.27	10.05
Model	Params	Dataset		KL	CLAP		
AudioLDM (Liu et al. 2023)	739 M	Multiple sources		1.76	0.43		
Stable Audio (Evans et al. 2024)	1B	AudioSparx		0.80	0.46		
Tango 2 (Majumder et al. 2024)	866 M	Audio-alpaca		1.12	0.57		

The table reports representative parameter counts and performance metrics, including PESQ, CSIG, CBAK, COVL, and SSNR for speech enhancement, and KL divergence and CLAP score for audio generation. The listed results are approximate and serve only as indicative comparisons, since the models were trained and evaluated under different conditions and datasets

high-resolution video synthesis. Generative Adversarial Networks (GANs) that incorporate U-Net, use its encoder–decoder structure to refine video generation, where the generator synthesizes realistic frames while the discriminator ensures temporal consistency. Meanwhile, diffusion models extend U-Net to process spatiotemporal data with a forward and backward denoising process, improving the generation of coherent and high-quality videos. These models, regardless of their specific framework, benefit from U-Net's ability to propagate fine-grained spatial features and ensure stability across multiple frames, enabling a range of tasks including video synthesis, inpainting, and stylized generation.

Diffusion models have revolutionized video generation by progressively refining noisy data to generate realistic outputs. In text-to-video frameworks, diffusion models (An et al. 2023; Bar-Tal et al. 2024; Zhou et al. 2022, 2024) apply U-Net-based architectures to enhance video synthesis, leveraging noise addition and denoising processes. These models leverage the inherent strengths of U-Net, including hierarchical feature extraction and skip connections, to ensure high-quality, temporally consistent video generation. By operating in a compressed latent space, these diffusion models significantly reduce computational costs while maintaining fidelity in the generation of long-form videos. The use of temporal attention and other spatiotemporal mechanisms further enhances the model's ability to generate coherent and contextually accurate motion across frames.

Other U-Net models have demonstrated significant advancements in video generation tasks, particularly in segmentation and synthesis. Motion U-Net (Rahmon et al. 2021) employs residual connections within convolutional layers to enhance motion segmentation, effectively capturing dynamic changes across video frames. AGUnet (Yin et al. 2021) introduces a dual-encoder architecture that processes both the current frame and a reference frame, facilitating fast one-shot video object segmentation and enabling efficient adaptation to new scenes. In the realm of video synthesis, AP-GAN (Zhang et al. 2022) utilizes a U-Net-based generator with dual encoders to preserve facial attributes during video face swapping, ensuring realistic and consistent facial features across frames. These models utilize U-Net's hierarchical structure and skip connections to maintain spatial coherence and temporal consistency, contributing to advancements in video generation tasks that do not rely on diffusion processes.

For video generation models, the parameter sizes are generally comparable to or slightly larger than their image-based counterparts, reflecting their architectural extension from image generation frameworks. These models typically adapt the U-Net backbone to han-

dle temporal consistency and frame-level coherence. The most commonly used datasets for evaluation include MSR-VTT and UCF-101, which provide diverse video samples for training and benchmarking. Similar to image generation, evaluation metrics such as PSNR, SSIM, and FID are widely applied, although Fréchet Video Distance (FVD) is more specifically employed to assess temporal quality and motion smoothness across generated frames. Representative models, their parameters, and corresponding FVD scores are summarized in Table 4.

4.5 3D generation

U-Net architectures have been effectively adapted for 3D data generation across various model types, including standard U-Net models, GANs, and diffusion models. These models leverage U-Net's hierarchical encoder–decoder structure and skip connections to capture spatial hierarchies and preserve fine-grained details in volumetric data. In GAN-based frameworks, U-Net serves as a powerful generator, synthesizing realistic 3D volumes by learning complex distributions from training data. Diffusion models, on the other hand, utilize U-Net to progressively refine noisy 3D data, enhancing the quality and realism of generated volumes. The adaptability of U-Net to 3D data, combined with its robust feature extraction capabilities, makes it a cornerstone in the development of models for tasks such as medical image segmentation, volumetric synthesis, and 3D reconstruction.

Several non-diffusion U-Net-based models have been developed to address specific challenges in 3D data processing. The original 3D U-Net (Çiçek et al. 2016) introduced a volumetric segmentation approach that learns dense segmentations from sparse annotations, making it suitable for applications with limited labeled data. dResU-Net (Raza et al. 2022) enhances this by incorporating deep residual learning to improve brain tumor segmentation from multimodal MRI scans, effectively capturing complex anatomical structures. S3D-UNet (Chen et al. 2018) further optimizes 3D U-Net by employing separable 3D convolutions, reducing computational complexity while maintaining high segmentation accuracy for brain tumors. In the realm of generative modeling, 3DGAUnet (Shi et al. 2023) integrates a 3D U-Net-based generator within a GAN framework to synthesize realistic 3D CT images of pancreatic ductal adenocarcinoma (PDAC) tumors, addressing data scarcity and enhancing diagnostic capabilities. These models demonstrate the versatility of U-Net architectures in handling 3D data across various domains, from segmentation to generative tasks.

Diffusion-based models have significantly advanced 3D generation from single images, leveraging pretrained 2D diffusion priors to synthesize detailed 3D structures. DreamFusion (Poole et al. 2022) employs a pretrained 2D text-to-image diffusion model to optimize a Neural Radiance Field (NeRF) using a novel loss function, enabling text-to-3D synthesis without requiring 3D training data. Video3D (Ha 2024) utilizes a video diffusion model and volume rendering techniques to generate 3D structures from a single image, enhancing the realism and coherence of the resulting 3D scenes. Efficient-3DiM (Jiang et al. 2023) introduces strategies to expedite training, reducing the time from 10 days to under 1 day, and demonstrates the capability to efficiently synthesize novel views from a single image. These models exemplify the potential of diffusion-based approaches in generating high-quality 3D content from limited 2D inputs.

For 3D generative models, the parameter sizes also follow the same trend as image-based U-Nets but are adapted for volumetric or multi-view data. Most 3D applications are

concentrated in medical imaging, where precise spatial representation and voxel-level accuracy are critical. The BraTS dataset is among the most frequently used benchmarks for this domain, providing standardized volumetric MRI scans for tumor segmentation and reconstruction tasks. Common metrics for evaluating 3D model performance include Intersection over Union (IoU) and the Dice Similarity Coefficient (DICE), which measure the overlap and consistency between predicted and ground-truth volumes. Similar to the video domain, these 3D models are summarized in Table 4.

4.6 Pose and action generation

U-Net architectures have been effectively employed in human pose and action generation tasks, including pose-guided synthesis, 3D pose estimation, and face completion across varying viewpoints. These tasks necessitate the model to preserve spatial structure while enabling flexibility in motion or articulation, a challenge that U-Net's encoder–decoder design addresses adeptly. In pose-guided person generation, U-Net-based architectures facilitate the synthesis of images of individuals in novel poses by integrating pose information with visual content. Similarly, in 3D pose estimation, U-Net's hierarchical feature extraction facilitates the capture of complex spatial relationships, enhancing the accuracy of pose predictions. Additionally, U-Net models have been utilized in face completion tasks, where they reconstruct occluded facial regions, improving recognition robustness across extreme head poses. Beyond human-centric applications, U-Net's flexibility extends to other domains, such as drone navigation, where lightweight U-Net models process visual inputs to generate segmentation maps that inform navigation decisions, effectively combining efficient image processing with strategic path planning.

Several models have demonstrated the versatility of U-Net architectures in pose and action generation tasks across diverse applications. The Pose Guided Person Generation Network (PG) (Ma et al. 2017) employs a two-stage U-Net-based architecture to synthesize images of individuals in novel poses. In the first stage, a coarse output is generated by fusing the source image and target pose using a U-Net-style generator. The second stage then refines the output to enhance realism and detail. Similarly, UNETR-Pose (tamasino52 2021) combines U-Net's hierarchical spatial feature learning with transformer-based global attention to achieve robust 3D pose estimation in multi-view setups, even in the presence of occlusion or ambiguous viewpoints. In face completion tasks, the Attention ResCUNet-GAN (Na et al. 2020) model uses a pair of residual U-Nets with attention mechanisms to complete occluded facial textures from partial observations, enhancing pose-invariant face recognition. Additionally, DreamPose (Karras et al. 2023) utilizes a diffusion-based method to generate animated fashion videos from still images, integrating body pose information with the input image to produce temporally consistent sequences. The HourGlass U-Net (Bulat et al. 2020) introduces gated skip connections and a hybrid HourGlass-U-Net architecture, which improves pose estimation accuracy and efficiency, demonstrating state-of-the-art results in the MPII and LSP datasets. Similarly, Variational U-Net (Esser et al. 2018) utilizes a conditional U-Net for shape-guided image generation, where appearance and shape are learned separately, enabling precise control over both aspects in pose-guided synthesis tasks. Finally, a lightweight U-Net model has been applied in drone navigation for obstacle avoidance, integrating reinforcement learning techniques to process visual inputs and generate segmentation maps that inform the drone's navigation decisions (Shin et al.

2020). These applications highlight the flexibility and effectiveness of U-Net-based architectures in generating high-quality outputs for both human pose and action generation, as well as non-human tasks such as drone navigation.

In pose and action generation tasks, U-Net architectures have been applied across a diverse range of applications, including human pose synthesis, facial expression generation, drone motion prediction, and robotic action planning. Because of this diversity, datasets used in these studies vary widely. Common benchmarks for human-centric tasks include Human3.6M, DeepFashion, and COCO-Pose, which contain annotated body keypoints or pose sequences. In contrast, motion- or control-oriented studies may rely on domain-specific datasets such as drone flight logs or simulated robotic environments. Evaluation metrics are equally diverse and depend heavily on the task; common measures include Mean Per Joint Position Error (MPJPE), Percentage of Correct Keypoints (PCK), Intersection over Union (IoU), and trajectory accuracy for motion tasks. However, it is difficult to make direct quantitative comparisons across studies, as each focuses on distinct modalities, motion types, and evaluation goals.

5 Limitations of U-Net in generative AI

Despite its foundational role in modern generative modeling, U-Net exhibits several inherent limitations that constrain its applicability in specific scenarios. These limitations are closely tied to the architectural principles that make U-Net effective in spatial feature extraction, yet they also define the boundaries of where alternative or hybrid approaches may be more suitable. Understanding these limitations is essential for guiding future research and informing decisions about when U-Net should or should not be deployed in generative systems.

5.1 Limited global context and long-range dependency modeling

One of the most fundamental limitations of U-Net is its restricted capacity to model long-range dependencies and global context. Its convolution-based encoder–decoder structure excels at capturing local spatial features but struggles to integrate information over large spatial or temporal extents. This shortcoming becomes evident in tasks such as text-to-image synthesis or video generation, where capturing complex semantic relationships or temporal consistency is crucial. Transformer-based architectures, with their self-attention mechanisms (Vaswani et al. 2017), or state space models like Mamba (Gu and Dao 2023), offer superior performance in these contexts by explicitly modeling global interactions. As a result, while U-Net remains a strong backbone for localized feature learning, it is less effective in scenarios requiring holistic context modeling or fine-grained cross-modal alignment.

5.2 Scalability and computational overhead

U-Net also faces scalability challenges, particularly in high-resolution or large-scale generative tasks. As input resolution increases, memory usage grows substantially, and the fixed-size receptive field of convolutional layers covers a proportionally smaller region of the data. This limits the network's capacity to extract meaningful global features and significantly increases computational requirements (Bar-Tal et al. 2024; Jiang et al. 2023). Such

limitations hinder the deployment of U-Net-based generative models in real-time systems or on resource-constrained devices, where efficient architectures or techniques like pruning (Cheng et al. 2023) and quantization (Nagel et al. 2021) may be necessary. In contrast, models designed with global context modeling or parameter efficiency in mind, such as efficient Transformers or SSM-based networks, often scale more gracefully under these conditions.

5.3 Limited semantic abstraction and cross-modal generalization

Another limitation arises from U-Net's reliance on pixel-level feature learning, which can constrain its ability to handle tasks requiring high-level semantic reasoning. While effective for spatial reconstruction tasks, U-Net often struggles in cross-modal generative applications—such as text-to-image (Radford et al. 2021) or audio-to-video synthesis—where capturing complex semantic relationships is essential. Without additional modules, such as cross-attention layers or language-conditioned embeddings, U-Net may fail to bridge the modality gap or generalize to out-of-domain data. Hybrid approaches, such as the U-Net Encapsulated Transformer (UET) (Ignacio et al. 2024; Wang et al. 2024), address some of these issues by combining convolutional spatial modeling with Transformer-based semantic understanding, yet the limitation remains inherent to U-Net's original design.

5.4 Interpretability and transparency

Interpretability remains a persistent challenge for U-Net-based generative models. Like most deep neural architectures, U-Net operates as a "black box," providing limited insight into how internal feature representations influence final outputs. This opacity is particularly concerning in safety-critical domains, such as healthcare (Abrantes and Rouzrokh 2024; Farahani et al. 2022; Velden et al. 2021) or autonomous systems (Omeiza et al. 2021; Sakai and Nagai 2021; Yeong et al. 2025), where understanding the reasoning behind model outputs is essential for trust and accountability. While explainable AI techniques—such as feature attribution (Kaur et al. 2022; Sharma et al. 2022) or influence tracing (Chowdhury and Srivastava 2024; Nguyen et al. 2021) can improve transparency, the lack of built-in interpretability mechanisms remains a fundamental limitation of the U-Net architecture.

5.5 Constraints in sequential and symbolic tasks

Finally, U-Net's convolutional nature limits its effectiveness in tasks that rely on sequential dependency modeling or symbolic reasoning. Autoregressive models and Transformer-based architectures are inherently better suited for such applications, as they are designed to capture token dependencies and temporal structures over extended sequences. While U-Net variants have been explored for tasks like text generation or decision-making, their performance often lags behind attention-based approaches (Chen et al. 2021), especially in natural language processing or reinforcement learning contexts. This limitation suggests that U-Net may not be the optimal choice in domains where modeling order, sequence, or symbolic structure is critical.

In summary, while U-Net remains a cornerstone of generative modeling due to its architectural simplicity and spatial feature extraction capabilities, its limitations in global context modeling, scalability, semantic abstraction, interpretability, and sequential reasoning

constrain its use in certain domains. Addressing these constraints through hybrid models, architectural innovations, and integration with complementary paradigms will be crucial to extending U-Net's role in next-generation generative AI systems.

6 Advantages of U-Net in generative AI

U-Net has emerged as a foundational architecture in Generative AI due to its encoder–decoder structure and skip connections, which enable effective feature extraction, preservation of fine details, and computational efficiency. Its hierarchical processing and multi-scale representations provide an ideal backbone for modeling both local and global features, which are essential for generating high-fidelity, context-aware outputs across modalities. This section highlights the core architectural benefits of U-Net and illustrates how they contribute to robust generative modeling in diverse domains.

6.1 Effective feature extraction

U-Net's encoder–decoder architecture enables powerful hierarchical feature extraction, which is crucial for tasks that require both global context and local details. This ability to capture multi-scale representations is fundamental for maintaining high output quality across diverse generative tasks. Through its downsampling and upsampling process, U-Net extracts both coarse and fine features, ensuring that the model can synthesize coherent outputs at various levels of abstraction.

In image generation, U-Net's hierarchical processing is crucial for preserving texture, shape, and composition, particularly in diffusion models such as Stable Diffusion and Imagen. Here, U-Net refines noisy latent representations, preserving high-fidelity visual details. Similarly, in image-to-image translation, U-Net's skip connections help preserve spatial coherence during tasks such as Pix2Pix, enabling more accurate sketch-to-photo conversion and domain adaptation. The PatchGAN discriminator further leverages U-Net's structure to enhance fine-grained texture analysis, focusing on localized details to improve classification accuracy. In PixelCNN++, the architectural similarities to U-Net enable the model to capture fine details across multiple resolutions, improving image generation performance. In video generation, U-Net's extension into spatiotemporal processing is essential for modeling motion dynamics across multiple frames. Models like Lumiere and MagicVideo leverage U-Net to ensure both temporal consistency and spatial coherence. These architectures use U-Net to process entire video sequences in a single pass, enhancing global temporal consistency and ensuring smooth transitions across frames. Upscale-A-Video further integrates 3D convolutions and temporal attention, improving the stability and quality of generated sequences.

U-Net architectures have been used in speech and audio generation to model both long-term dependencies and local waveforms. Wave-U-Net excels at speech enhancement by capturing large temporal contexts in the encoder while reconstructing fine acoustic details in the decoder. U-Net's application to source separation and voice conversion demonstrates its ability to preserve important temporal structures while disentangling speaker characteristics from content. Additionally, UNetGAN, which integrates U-Net with GANs, enhances

performance in extremely low SNR conditions, showcasing the model's flexibility in adversarial training setups.

In 3D volumetric generation, U-Net's hierarchical feature extraction plays a crucial role in interpreting and reconstructing spatial structures from sparse data. 3D U-Net and Efficient-3DiM apply this principle to medical imaging and 3D object generation, enabling accurate volumetric reconstructions. Similarly, Video3D uses a diffusion-based U-Net to generate temporally coherent multi-view 3D representations from a single image. In pose and action generation, U-Net's encoder–decoder design supports hierarchical learning of spatial features. Models like PG and UNETR-Pose utilize this structure to generate human poses and actions, ensuring spatial alignment and consistency across various poses and viewpoints. In facial image completion, U-Net-based models, such as Attention ResCUNet-GAN, utilize hierarchical feature extraction to infer missing facial regions and enhance pose-invariant face recognition.

Through its ability to learn hierarchical representations and preserve fine details across diverse modalities, U-Net remains a cornerstone in generative AI, enabling deep feature extraction and enhancing performance in both common and emerging applications.

6.2 Preservation of fine details

A defining feature of U-Net is its exceptional ability to preserve fine-grained details—whether spatial, temporal, or structural—across generative tasks. This capability is primarily due to U-Net's encoder–decoder architecture with skip connections, which allow low-level features to bypass the bottleneck and contribute directly to the reconstruction process. These connections mitigate the loss of resolution and high-frequency features, which are common in deep architectures, ensuring that high-quality, detail-preserving outputs are generated.

In image generation, diffusion models like Stable Diffusion and Imagen leverage U-Net's denoising function to refine images, preserving texture and edge details progressively. In image-to-image translation, U-Net-based models, such as Pix2Pix and PatchGAN, rely on skip connections to maintain spatial coherence and fine-grained features during tasks like sketch-to-photo conversion and domain adaptation. Similarly, in PixelCNN++, which shares architectural similarities with U-Net, downsampling and shortcut connections enhance the model's ability to capture fine details at multiple resolutions, improving overall image generation performance. In video generation, Lumiere and MagicVideo utilize U-Net to maintain both spatial and temporal coherence, resulting in high-resolution, temporally consistent video outputs. In particular, Upscale-A-Video integrates 3D convolutions and temporal attention to reduce flickering and enhance stability, ensuring smooth transitions and maintaining fine details across frames.

In speech and audio generation, Wave-U-Net and Attention Wave-U-Net showcase U-Net's ability to model both global temporal structures and fine acoustic details, making them effective in speech enhancement and denoising tasks. Similarly, UNetGAN enhances performance in low SNR conditions by integrating U-Net with dilated convolutions, which capture large temporal contexts, and ensures high-quality audio output. For 3D generation, U-Net's hierarchical feature extraction supports the restoration of volumetric consistency, as seen in 3DGAUnet and Efficient-3DiM, where U-Net helps refine noisy or incomplete volumetric data, producing accurate 3D reconstructions for medical imaging and object generation. Finally, in pose and action generation, models such as PG and UNETR-Pose

utilize U-Net's hierarchical processing to preserve spatial features, ensuring the accurate generation of human poses across varying viewpoints or frames. The Attention ResCUNet-GAN model utilizes U-Net with attention mechanisms to complete missing facial textures, improving pose-invariant face recognition even under extreme head poses.

Across modalities, U-Net's encoder–decoder structure and skip connections ensure that intricate spatial, temporal, and structural details are preserved, making it a powerful architecture for high-fidelity generative tasks that require fine-grained accuracy and detail. Figure 6 visualizes the role of skip connections in preservation of fine details.

6.3 Computational efficiency

U-Net's computational efficiency is one of its standout features, enabling high-quality generation across multiple modalities with reduced resource consumption. The key to U-Net's efficiency lies in its convolutional design, hierarchical processing, and support for parallelization, which makes it particularly attractive for large-scale and resource-constrained applications in generative AI. Figure 7 shows the advantage of U-Net compared to Autoregressive and Transformer-based models.

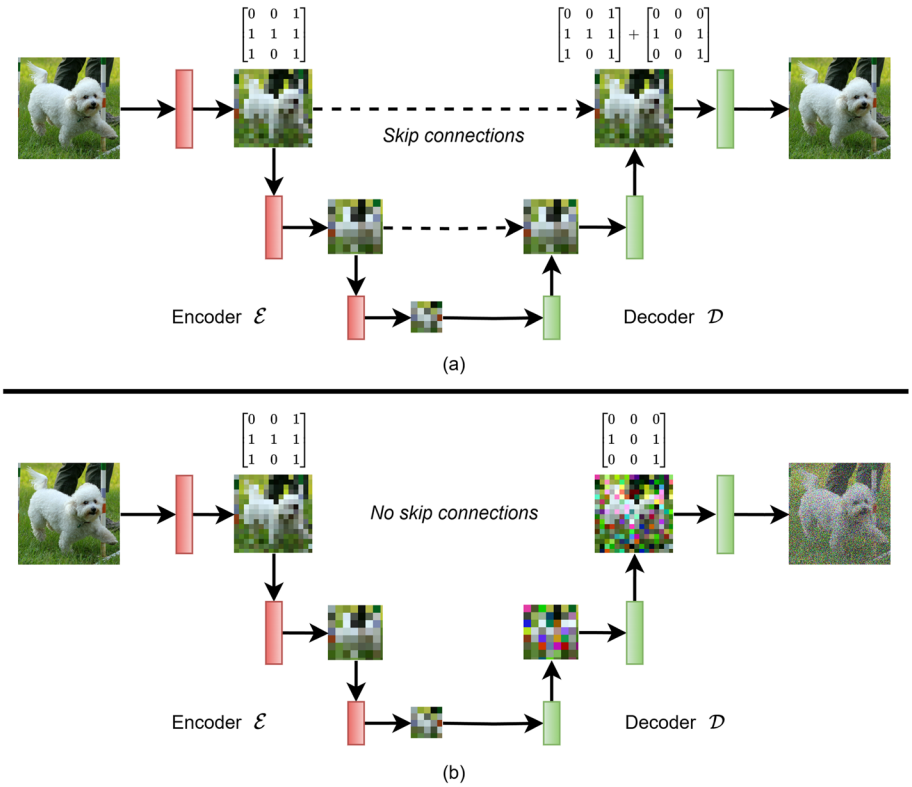

Fig. 6 Role of skip connections in U-Net. **a** Standard U-Net, where the encoder captures multi-scale features and the decoder reconstructs outputs using skip connections to transfer spatial details. **b** Standard encoder–decoder network without skip connections, which leads to blurrier outputs and loss of high-frequency details. The example image is from the Flickr8k dataset (Hodosh et al. 2013)

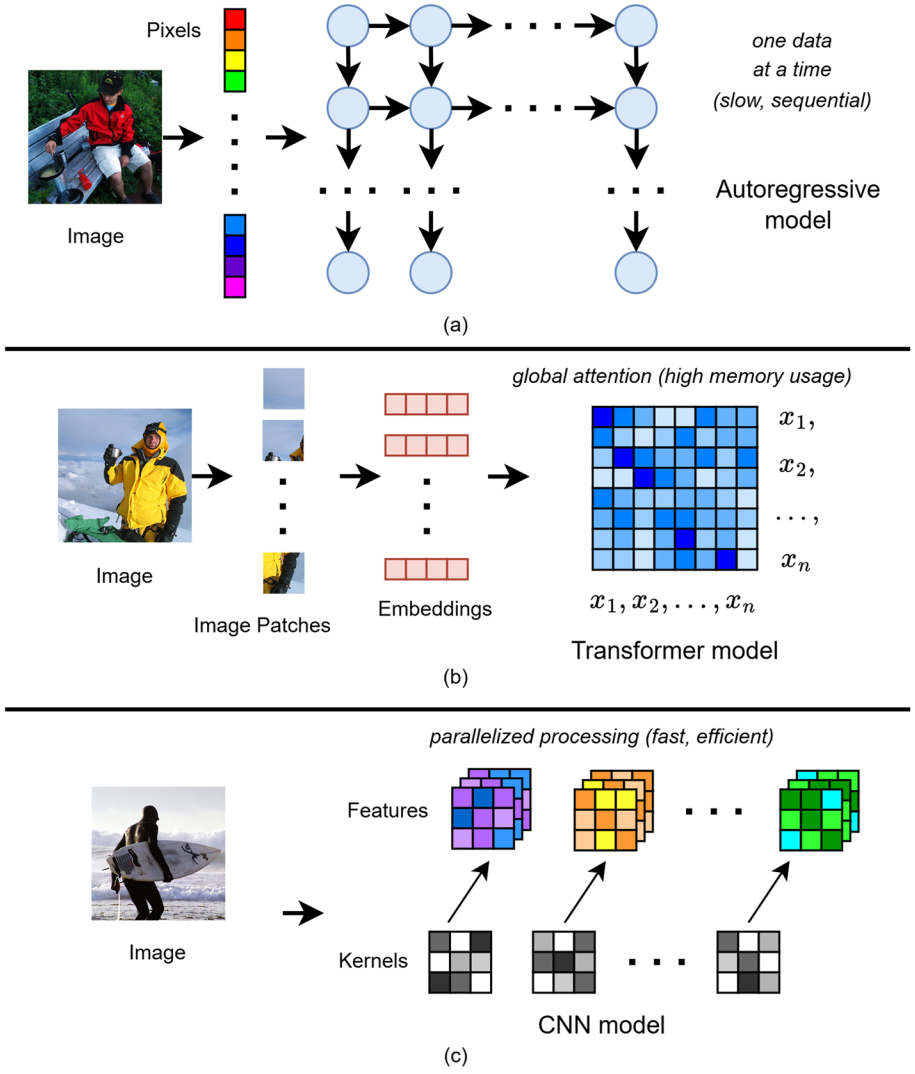

Fig. 7 Comparison of generative modeling approaches. **a** Autoregressive model generating pixels sequentially, resulting in slow processing. **b** Transformer-based model capturing long-range dependencies via global attention but with high memory cost. **c** U-Net convolutional architecture enabling fast, parallel image generation with lower computational overhead. Sample images are from the Flickr8k dataset (Hodosh et al. 2013)

Unlike autoregressive models, which generate outputs sequentially and suffer from slow inference speeds, U-Net supports parallelized processing across entire inputs. This characteristic is especially beneficial for image and audio generation, where high-resolution outputs must be produced rapidly. For example, in Stable Diffusion, U-Net processes entire images during the denoising steps, dramatically improving inference speed compared to pixel-by-pixel generation. Similarly, in Wave-U-Net, the U-Net model operates directly in

the time domain, bypassing the need for computationally expensive spectrograms in speech enhancement tasks, which accelerates signal processing while maintaining high quality.

Compared to transformer-based models, which typically exhibit quadratic time complexity due to their global attention mechanisms, U-Net leverages localized convolutional operations that scale linearly with input size. This linear scaling makes U-Net more memory-efficient, especially when working with high-resolution inputs where global attention becomes prohibitively expensive. For example, in UET, U-Net reduces the embedding dimension before passing the compressed representations to the Transformer block, significantly reducing the number of parameters and memory usage, making it more suitable for efficient language modeling (Ignacio et al. 2025).

In 3D generation, models like Efficient-3DiM showcase how U-Net's efficient design extends to volumetric data. The model reduces computational overhead while generating novel views of 3D shapes, utilizing U-Net's denoising backbone within a diffusion framework to improve efficiency. Furthermore, U-Net's multi-scale processing enhances computational efficiency across domains. In video generation, Lumiere and MagicVideo employ U-Net's hierarchical structure to process spatiotemporal data in a computationally efficient manner, enabling high-quality video synthesis while managing temporal dependencies across frames. The Upscale-A-Video framework incorporates 3D convolutions and temporal attention modules to reduce flickering and ensure stability during video generation, further improving spatial and temporal coherence. In pose generation, models like PG and UNETR-Pose combine U-Net's hierarchical feature extraction with other architectures, enabling robust generation with fewer parameters and reduced computational costs in tasks such as 3D pose estimation and human figure synthesis.

By maintaining its hierarchical design, U-Net supports efficient feature extraction and generation across various applications. Its ability to balance computational efficiency with high-quality output makes it an ideal backbone for scalable, real-time, and multimodal generative systems. It ensures that these models remain practical even in resource-constrained environments.

6.4 Additional advantages

Beyond its core strengths in feature extraction, detail preservation, and computational efficiency, U-Net offers several other notable advantages that bolster its adaptability across diverse, multimodal generative tasks. These additional benefits, including robustness to noisy inputs, flexibility across modalities, and seamless integration with modern techniques, have contributed significantly to the widespread use of U-Net in generative AI.

Robustness to Noisy Inputs U-Net's design, originally intended for medical image segmentation, gives it a remarkable resilience against noisy or incomplete data, such as occlusions, artifacts, and degraded signals. This quality proves essential in generative tasks where input quality may vary. For instance, Stable Diffusion and Imagen rely on U-Net to denoise images progressively, generating stable outputs even from highly corrupted inputs. Additionally, models like Wave-U-Net have demonstrated the efficacy of U-Net in recovering clean speech from noisy audio signals, improving signal quality even under challenging conditions. This robustness allows U-Net-based models to excel in real-world generative tasks, where data quality cannot always be controlled.

Flexibility Across Modalities U-Net's architecture is highly versatile, making it well-suited for various data modalities, including images, audio, video, 3D structures, and human pose and action generation. For example, Pix2Pix and PatchGAN utilize U-Net for image-to-image translation, where U-Net's skip connections preserve fine-grained details during translation, thereby improving the quality and consistency of the generated images. Similarly, in audio tasks, models like Wave-U-Net operate directly on raw audio waveforms, ensuring that both long-term and local temporal dependencies are captured during speech enhancement. In 3D generation, Video3D and Efficient-3DiM use U-Net to generate temporally consistent multi-view 3D representations from sparse data, highlighting U-Net's ability to handle complex volumetric and spatiotemporal data. This flexibility makes U-Net an invaluable backbone in multimodal generative pipelines, supporting tasks from pose and action generation to voice conversion.

Seamless Integration with Modern Techniques U-Net's modular structure facilitates easy integration with cutting-edge techniques, such as attention mechanisms, residual blocks, and Transformer modules, enabling it to stay at the forefront of generative AI research. For example, UET combines U-Net's dimensionality reduction with Transformer efficiency, enhancing memory usage and computational speed in language modeling. Moreover, U-Net can seamlessly integrate self-attention in models like Attention Wave-U-Net, which improves the ability to process raw waveforms directly while suppressing background noise for superior speech enhancement. Similarly, Lumiere and MagicVideo incorporate U-Net within diffusion models to generate temporally consistent and high-quality video content, while models like Upscale-A-Video use temporal attention to ensure stability and coherence in video synthesis. These integrations allow U-Net to adapt to the latest advancements in deep learning, ensuring it remains a highly competitive and forward-compatible architecture.

In summary, U-Net's robustness, multimodal flexibility, and modular extensibility make it not only a powerful generative tool but also a sustainable and adaptable architecture. Its ability to integrate with modern deep learning techniques ensures its continued relevance and effectiveness in a broad array of generative AI applications.

7 Challenges in U-Net for generative AI

7.1 Loss of spatial information

Despite U-Net's significant contributions to generative AI, several inherent challenges arise when applying its architecture to complex, high-dimensional generative tasks. One of the primary challenges is the *loss of spatial information*. While U-Net's encoder–decoder structure, bolstered by skip connections, effectively retains low-level spatial features during the downsampling and upsampling processes, the repeated downsampling in the contracting path still leads to the loss of fine-grained details. This issue becomes particularly pronounced in high-resolution generative tasks such as image synthesis, video generation, and 3D reconstruction, where precise pixel-level information is critical. As the architecture deepens, the ability to preserve fine details diminishes, which can limit the model's effectiveness in tasks that demand high-resolution outputs or intricate texture and structure synthesis.

7.2 Computational cost

Another significant hurdle is *computational cost*. U-Net's architecture, particularly when adapted for large datasets or high-resolution data, can be extremely memory-intensive. The combination of deep encoder–decoder stacks and skip connections demands substantial computational resources, which becomes increasingly problematic in generative tasks dealing with high-dimensional data, such as video generation or 3D volumetric reconstruction. For instance, in models like Lumiere (Bar-Tal et al. 2024) and Efficient-3DiM (Jiang et al. 2023), the ability to generate temporally coherent video or volumetric data requires the processing of a large number of frames or volumetric slices, further compounding the memory and computational demands. These limitations are especially challenging for deployment in resource-constrained environments, where there may be insufficient GPU memory or computational power to process such high-dimensional data in real-time or at scale. Figure 8 illustrates the computational cost challenges associated with the U-Net architecture.

7.3 Generalization

Generalization issues present another challenge for U-Net-based models, particularly when applied to more diverse or complex data modalities. U-Net is inherently designed for pixel-level transformations and performs best when operating on data that is similar to what it was trained on. However, its performance can degrade when applied to out-of-domain data or tasks with significant variations not seen during training. For example, in multi-modal tasks like text-to-image generation, U-Net's reliance on pixel-level features may make it difficult to capture the intricate relationships between different modalities (e.g., between textual descriptions and image features) without further enhancements. Integrating U-Net with architectures like Transformers or cross-attention mechanisms, such as in UET (Ignacio et al. 2024; Wang et al. 2024), can help bridge this gap by capturing long-range dependencies and improving the model's ability to generalize to unseen data or more complex multi-modal tasks.

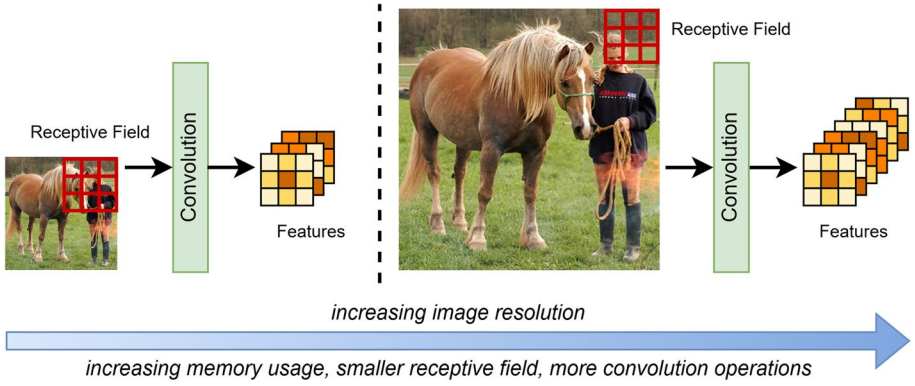

Fig. 8 Challenges of U-Net architectures with increasing image resolution. As resolution grows, memory consumption increases, receptive fields become proportionally smaller, and more convolution operations are required to capture meaningful features. This trade-off highlights a key limitation of convolution-based models, where scaling to high-resolution inputs demands deeper architectures or larger filters. Example images are from the Flickr8k dataset (Hodosh et al. 2013)

7.4 Interpretability

Additionally, the *interpretability* of U-Net-based models remains a concern. Like many deep learning architectures, U-Net operates as a "black-box", making it difficult to discern how or why a model produces a specific output. This lack of transparency is particularly problematic in domains where understanding the model's decision-making process is essential, such as in medical imaging (Abrantes and Rouzrokh 2024; Farahani et al. 2022; Velden et al. 2021) or autonomous systems (Omeiza et al. 2021; Sakai and Nagai 2021; Yeong et al. 2025). Moreover, the user interface and experience (UI/UX) design for generative AI plays a critical role in facilitating human understanding of the model's behavior and predictions. Proper UI/UX for these systems can bridge the gap between the black-box nature of the model and user comprehension, making the decision-making process more transparent and trustworthy (Kim et al. 2024). The ability to trace back through the layers and understand feature extraction, decision-making, and the overall generative process could significantly enhance the trustworthiness of U-Net models in critical applications.

7.5 Adaptation for generative tasks

Lastly, *adaptation for generative tasks* poses its own set of challenges. While U-Net excels at pixel-level transformations, such as in segmentation or inpainting, its direct application to more complex generative tasks that require higher-level abstraction or control can be challenging. For instance, in tasks like text-to-image generation, U-Net alone may not have the capacity to model complex semantic structures without integrating additional modules such as Transformers or adversarial networks. Hybrid architectures, such as UET, often offer more sophisticated generative capabilities by incorporating additional components, including latent space manipulation, adversarial loss functions, or temporal modeling. This shows that while U-Net provides a solid foundation, achieving advanced generative results requires combining it with other architectures that can handle higher-level abstractions and modality-specific requirements.

In conclusion, while U-Net remains a powerful backbone for many generative models, addressing challenges such as loss of spatial information, high computational demands, generalization to diverse tasks, interpretability, and adaptation for complex generative tasks is crucial for expanding its applicability. Advances in hybrid architectures, integration with attention mechanisms, and the development of more efficient versions of U-Net are necessary to overcome these challenges and further solidify its role in the future of generative AI.

8 Future directions for U-Net in generative AI

The future of U-Net in generative AI will likely focus on addressing its current limitations while building on its strengths. U-Net has shown considerable potential across a range of generative tasks. However, as applications continue to expand into new modalities, ongoing research is necessary to enhance their adaptability and performance.

8.1 Enhanced architectures

A key area of research lies in developing more efficient and robust U-Net variants. While U-Net's hierarchical encoder–decoder structure and skip connections have contributed to its success in generative tasks, there is still room for improvement in its ability to capture long-range dependencies and contextual information. Emerging attention mechanisms, such as Global Spatial-Channel Attention (Song et al. 2021), could further bolster U-Net's ability to focus on relevant regions within the input and capture dependencies across large spatial areas. Additionally, incorporating more efficient convolutional blocks such as depthwise separable convolutions (Chollet 2016; Howard et al. 2017), pointwise group convolutions (Zhang et al. 2017), and techniques like compound scaling (Tan and Le 2019) and parameter reduction (Iandola et al. 2016) will be crucial to enable real-time applications, particularly in resource-constrained environments. These advances will be essential for high-resolution tasks, such as 3D image generation or text-to-image synthesis, where maintaining output quality without sacrificing efficiency remains a challenge.

8.2 Integration with other generative models

The future of U-Net in generative AI will also see deeper integration with other prominent generative models, such as Generative Adversarial Networks (GANs) and Diffusion models. While U-Net has already been successfully employed as a generator in GANs, its potential extends beyond this role, with opportunities to be incorporated more effectively into the adversarial training process or seamlessly integrated into the diffusion process. Such advancements could significantly enhance both generation quality and efficiency. For instance, in models like Stable Diffusion and Imagen, U-Net's role in the denoising process has already proven invaluable, and further research may expand its application to improve output diversity and temporal coherence, particularly in tasks like video generation.

Figure 9 illustrates the concept of a multi-modal UET framework for cross-modal learning and text-to-image generation. In this framework, UET processes text and image embeddings independently, aligning them through contrastive learning models such as CLIP (Radford et al. 2021), before generating an image from the text-based embedding, facilitating efficient text-to-image translation. Another promising direction for integrating U-Net into generative models is applying diffusion processes to text generation. In this context, UET can guide text generation through progressive denoising. Figure 10 illustrates this process, where text embeddings undergo forward and backward diffusion, with conditioning applied to influence the output generation, enabling controlled and high-quality text synthesis. Recent work has also explored foundation-scale multimodal models, such as GPT-4V, which demonstrate strong zero-shot visual recognition capabilities. For example, GPT-4V has been shown to identify vehicle brands directly from images without any task-specific fine-tuning (Kerdvibulvech 2025), highlighting complementary directions for integrating U-Net architectures with large vision-language models in future multimodal systems.

Recent multimodal diffusion architectures, such as FreeU (Si et al. 2023) and VLOGGER (Corona et al. 2024), have explored specific enhancements to the diffusion-based U-Net framework. FreeU improves generative quality by re-weighting skip and backbone features within the existing diffusion U-Net. In contrast, VLOGGER extends diffusion models with temporal conditioning for audio-driven human video synthesis. In contrast,

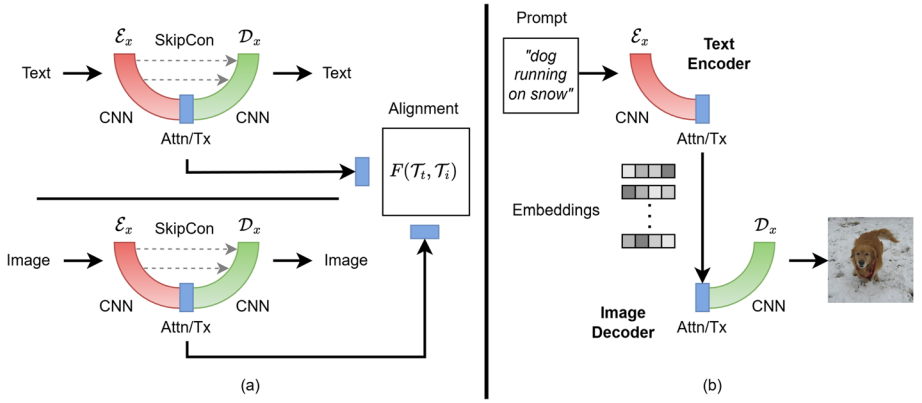

Fig. 9 Multi-modal UET framework for cross-modal learning and text-to-image generation. **a** Alignment stage, where separate UETs process text and image inputs and align their feature representations using a contrastive model such as CLIP, enabling a shared semantic space. **b** Text-to-image generation, where text embeddings encoded by UET guide an image decoder to synthesize visual outputs based on aligned textual representations. Example data are from the Flickr8k dataset (Hodosh et al. 2013)

the proposed U-Net Encapsulated Transformer (UET) represents a broader architectural innovation that integrates full Transformer modules within the U-Net body, rather than relying on peripheral attention or feature scaling mechanisms. This encapsulated design allows UET to inherit U-Net's spatial precision while enabling long-range contextual modeling similar to that of large language models. Consequently, UET can theoretically unify spatial, temporal, and semantic representations within a single generative framework, bridging the gap between visual and multimodal reasoning in future diffusion systems.

One promising avenue for future research is the potential combination of U-Net with state space models (SSMs) like Mamba (Gu and Dao 2023). Currently, Mamba-based models have been predominantly used in medical image segmentation tasks (Ma et al. 2024; Ruan and Xiang 2024; Wang et al. 2024; Xing et al. 2024), leveraging its ability to model long-range dependencies. Given that Mamba was initially designed as a competitor to Transformer models, and considering that UET is a variant of Transformer, it is feasible to integrate elements of Mamba, Transformer, U-Net, and even Diffusion within a unified framework. This combination could leverage the strengths of each model, enabling enhanced generative capabilities in complex multimodal tasks, including text-to-image generation.

8.3 Scalability and efficiency

A critical challenge moving forward is making U-Net models scalable and efficient, particularly in the context of high-resolution data and 3D applications. Techniques like network pruning (Cheng et al. 2023), quantization (Nagel et al. 2021), and distributed training (Zeng et al. 2023) will play a vital role in addressing the computational demands of U-Net when applied to larger datasets or more complex modalities. In particular, the application of U-Net in volumetric data generation, such as 3D medical imaging or point cloud processing, will necessitate innovations to ensure that memory usage and computational cost remain manageable without compromising output quality. Efficient training protocols will

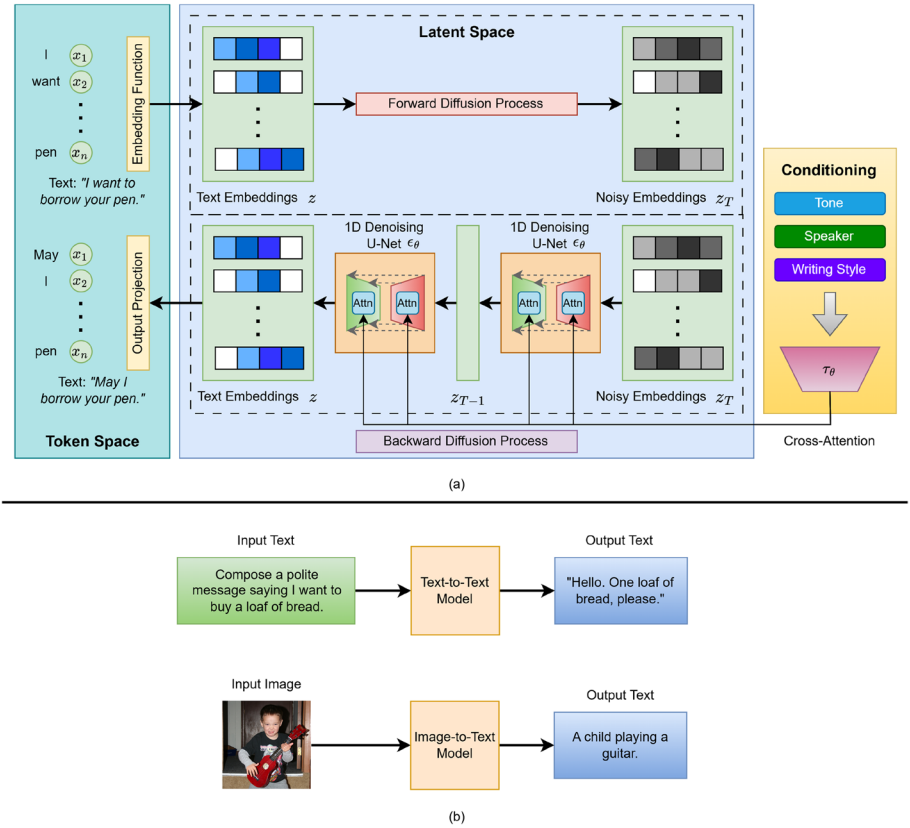

Fig. 10 Diffusion-based text generation using a 1D U-Net. **a** Forward diffusion adds noise to token embeddings, which are conditioned on additional information such as topic or writing style through a conditioning encoder and cross-attention, and the backward process refines them to generate coherent text. **b** Example use cases include text-to-text and image-to-text generation. The image example is from the Flickr8k dataset (Hodosh et al. 2013)

be essential for enabling U-Net to handle these expansive tasks while maintaining its status as an efficient and effective architecture.

8.4 Improved generalization and robustness

As generative models expand to diverse and complex datasets, U-Net's generalization abilities must be improved to handle variations in input data effectively. Current models may struggle with out-of-domain data, or data with significant variations not encountered during training. Future research will likely focus on enhancing U-Net's robustness through techniques such as domain adaptation (Chen et al. 2019; Redko et al. 2020) and meta-learning (Khoee et al. 2024; Li et al. 2018), which can help U-Net models generalize more effectively across various tasks and domains. This will be particularly important for applications in fields such as healthcare, where data can be highly heterogeneous, and robustness to unseen data is critical.

8.5 Explainable AI (XAI)

Another essential direction for future research is improving the interpretability of U-Net-based generative models. As U-Net is increasingly applied to sensitive fields, such as healthcare, where model decisions can have life-altering consequences, understanding how U-Net generates its outputs becomes crucial. Research into explainable AI (XAI) techniques, such as methods for visualizing the learned features of U-Net via heatmap (Kaur et al. 2022; Sharma et al. 2022) or tracing the influence of specific input components on the final output (Chowdhury and Srivastava 2024; Nguyen et al. 2021), will be critical. Providing transparent explanations for the model's decisions will help in high-stakes applications and ensure trust and reliability in U-Net-based generative models.

8.6 Novel applications

U-Net's versatility positions it well for exploration in emerging applications within generative AI. Having demonstrated success across tasks such as segmentation, inpainting, and various other generative applications, U-Net has proven adaptable across a broad spectrum of domains. Moving forward, one of the promising directions is its expanded use in multimodal generation, where U-Net could facilitate the synthesis of complex outputs from diverse data sources, including text, image, audio, and video. For example, U-Net can play a pivotal role in generative tasks such as text-to-3D or audio-to-video synthesis, where cross-modal transformations and feature fusion are crucial for creating coherent, high-quality results. As generative AI continues to push boundaries in both creative and industrial fields, U-Net's architecture is well-positioned to remain a central component, driving innovation in content creation across various modalities. Figure 11 illustrates a potential unified multimodal diffusion model, where both text and image modalities are generated through a consistent diffusion process. This model employs a Transformer architecture that is trained in a manner similar to a diffusion model, thereby integrating the strengths of both approaches to achieve high-quality text-to-image synthesis.

While U-Net has traditionally been employed in visual analysis tasks, its potential as a core component for decision-making frameworks remains largely unexplored. Inspired by the success of Decision Transformer (Chen et al. 2021) in reinforcement learning (RL) applications, one proposal is to extend UET to imitation learning and RL tasks. In this setup, UET could replace or augment the Transformer blocks in the Decision Transformer, leveraging its efficient encoding-decoding architecture to better capture temporal dependencies and structured behaviors. This adaptation would enable UET-based models to serve not only in visual perception but also in sequential decision-making environments such as autonomous driving, drone navigation, and robotic manipulation. By utilizing UET in this way, U-Net can evolve beyond static data generation and analysis, offering a lightweight and scalable backbone for learning control policies in complex, real-world scenarios.

In summary, the future of U-Net in generative AI lies in its ability to evolve through enhancements in architecture, integration with other generative models, and adaptability to increasingly complex and multi-modal tasks. By overcoming its current challenges and leveraging its inherent strengths, U-Net will continue to play a crucial role in shaping the future of generative modeling across various domains, ensuring that it remains at the forefront of the AI landscape.

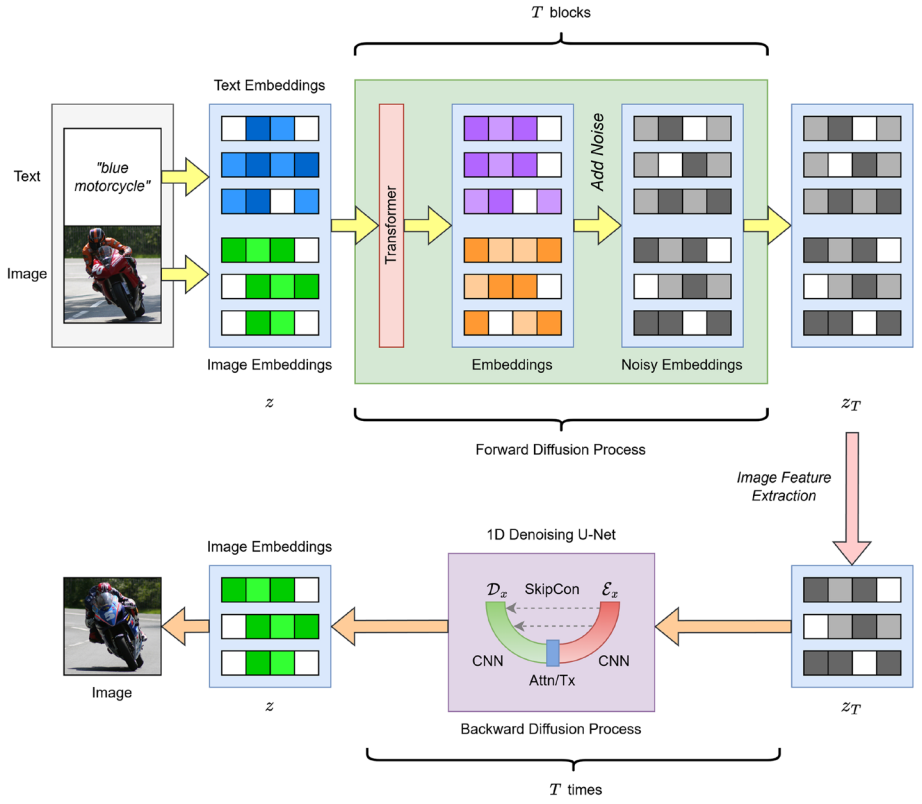

Fig. 11 Multimodal diffusion framework for text-to-image generation. Text and image inputs are transformed into embeddings and jointly processed through a Transformer to model contextual relationships. Progressive noise is added during forward diffusion, producing noisy embeddings that are refined by a 1D U-Net during backward diffusion to recover the image embedding. The final output is reconstructed into image form. Example data are from the Flickr8k dataset (Hodosh et al. 2013)

9 Conclusion

This survey systematically examined the progressive development of the U-Net architecture, tracing its evolution from its initial application in biomedical image segmentation to its pivotal position within contemporary generative AI systems. The integration of key architectural enhancements, such as attention mechanisms, residual connections, and optimized normalization techniques, has established the U-Net as a highly versatile backbone. This foundation significantly contributes to enhanced training stability, increased model expressiveness, and superior synthesis fidelity across a diverse array of generative paradigms.

We reviewed U-Net's integration in diffusion, GAN, and autoregressive models, where it enables high-resolution generation, structured image-to-image translation, and improved local feature modeling. Its encoder–decoder structure and skip connections also underpin its adaptability across multiple modalities, including image, audio, video, 3D, and multimodal tasks, highlighting its scalability and modular design.

Despite its success, U-Net faces ongoing limitations, including high computational demands, limited global context modeling, and challenges in scaling to large or multimodal datasets. Addressing these issues will require lightweight architectures, hybrid convolution-attention designs, and improved cross-domain alignment techniques. Nevertheless, U-Net remains a cornerstone of generative modeling. As generative AI advances toward more complex modalities and applications, continued innovations around U-Net will ensure its enduring relevance in building efficient, scalable, and high-quality generative systems.

In conclusion, this analysis underscores the enduring significance of the U-Net as a foundational architectural paradigm central to modern generative artificial intelligence. By effectively bridging conventional convolutional design with emerging generative modeling techniques, the U-Net structure persistently furnishes the necessary structural and conceptual backbone for subsequent domain-agnostic innovation. This adaptability reaffirms its status as one of the most influential and versatile neural network architectures in the contemporary field of deep learning.

Acknowledgements The authors gratefully acknowledge the support of the Institute of Information & Communications Technology Planning & Evaluation, funded by the Ministry of Science and ICT, and the Basic Science Research Program through the National Research Foundation of Korea, funded by the Ministry of Education. Their generous funding and continued support made this work possible.

Author Contributions Marvin John Ignacio: Methodology, Formal analysis, Investigation, Visualization, Writing-Original Draft & Editing. Sangyun Shin: Methodology, Analysis, and Review. Hulin Jin: Software and Visualization. Seong Joon Yoo: Resources, Funding, and Review. Dongil Han: Analysis, Software, and Funding. Yong-Guk Kim: Conceptualization, Methodology, Supervision, Funding, Writing & Editing.

Funding This work was supported by the Information Technology Research Center (ITRC) support program (IITP-2025-RS-2022-00156354) and a grant funded by the Korean government Ministry of Science and Information Technology (MSIT) (No. RS-2019-II190231), as well as by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540). Also, this work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2025-25443732, Research on Ethical Reasoning and Metacognition for Human-Aligned AGI, 20%) and by the National Program for Excellence in SW, supervised by IITP in 2025 (2024-0-00037).

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

Abrantes J, Rouzrokh P (2024) Explaining explainability: the role of XAI in medical imaging. Eur J Radiol 173:111389

Al-hammuri K, Gebali F, Kanan A, Chelvan IT (2023) Vision transformer architecture and applications in digital health: a tutorial and survey. Vis Comput Ind Biomed Art 6(1):14

Ansar W, Goswami S, Chakrabarti A (2024) A survey on transformers in NLP with focus on efficiency. arXiv:2406.16893

An J, Zhang S, Yang H, Gupta S, Huang J-B, Luo J, Yin X (2023) Latent-shift: latent diffusion with temporal shift for efficient text-to-video generation. arXiv preprint arXiv:2304.08477
Armato A, Fanucci L, Scilingo EP, Rossi DD (2011) Low-error digital hardware implementation of artificial neuron activation functions and their derivative. Microprocess Microsyst 35:557–567
Asperti A, Evangelista D, Piccolomini EL (2021) A survey on variational autoencoders from a green ai perspective. SN Comput Sci 2(4):301
Ba JL, Kiros JR, Hinton GE (2016) Layer normalization. arXiv preprint arXiv:1607.06450
Bar-Tal O, Chefer H, Tov O, Herrmann C, Paiss R, Zada S, Ephrat A, Hur J, Liu G, Raj A, Li Y, Rubinstein M, Michaeli T, Wang O, Sun D, Dekel T, Mosseri I (2024) Lumiere: a space-time diffusion model for video generation. In: SIGGRAPH Asia 2024 conference papers, pp 1–11
Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M-T, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners. arXiv:2005.14165
Bulat A, Kossaifi J, Tzimiropoulos G, Pantic M (2020) Toward fast and accurate human pose estimation via soft-gated skip connections. In: 2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020), pp 8–15
Cao H, Tan C, Gao Z, Xu Y, Chen G, Heng P-A, Li SZ (2022) A survey on generative diffusion models. IEEE Trans Knowl Data Eng 36:2814–2830
Cao Y, Li S, Liu Y, Yan Z, Dai Y, Yu PS, Sun L (2023) A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT. arXiv:2303.04226
Chen L, Lu K, Rajeswaran A, Lee K, Grover A, Laskin M, Abbeel P, Srinivas A, Mordatch I (2021) Decision transformer: reinforcement learning via sequence modeling. Neural Inf Process Syst 34:15084–15097
Cheng H, Zhang M, Shi JQ (2023) A survey on deep neural network pruning: taxonomy, comparison, analysis, and recommendations. IEEE Trans Pattern Anal Mach Intell 46:10558–10578
Chen W, Liu B, Peng S, Sun J, Qiao X (2018) S3D-UNet: separable 3D U-Net for brain tumor segmentation. In: BrainLes@MICCAI
Chen Z, Zhuang J, Liang X, Lin L (2019) Blending-target domain adaptation by adversarial meta-adaptation networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2248–2257
Chollet F (2016) Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 1800–1807
Chowdhury P, Srivastava G (2024) Enhanced classification and segmentation of brain tumors in MRI images using custom CNN and U-Net models with XAI. In: International conference on pattern recognition. Springer, pp 1–16
Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016) 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 424–432
Corona E, Zanfir A, Bazavan EG, Kolotouros N, Alldieck T, Sminchisescu C (2024) Vlogger: multimodal diffusion for embodied avatar synthesis. In: 2025 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 15896–15908
DeepSeek-AI (2024) DeepSeek LLM: scaling open-source language models with longtermism. arXiv:2401.02954
Devlin J, Chang M-W, Lee K, Toutanova K (2019) Bert: pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics
Donahue D, Lialin V, Rumshisky A (2019) Injecting hierarchy with u-net transformers. arXiv preprint arXiv:1910.10488
Esser P, Sutter E, Ommer B (2018) A variational U-Net for conditional appearance and shape generation. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 8857–8866
Evans Z, Carr C, Taylor J, Hawley SH, Pons J (2024) Fast timing-conditioned latent audio diffusion. arXiv:2402.04825
Farahani FV, Fiok K, Lahijanian B, Karwowski W, Douglas PK (2022) Explainable AI: a review of applications to neuroimaging data. Front Neurosci 16:906290
Giri R, Isik U, Krishnaswamy A (2019) Attention wave-U-Net for speech enhancement. In: 2019 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA). IEEE, pp 249–253
Gozalo-Brizuela R, Garrido Merchan E (2024) A survey of generative AI applications. J Comput Sci 20(8):801–818. https://doi.org/10.3844/jcssp.2024.801.818
Gu A, Dao T (2023) Mamba: linear-time sequence modeling with selective state spaces. arXiv:2312.00752
Gul S, Khan MS (2023) A survey of audio enhancement algorithms for music, speech, bioacoustics, biomedical, industrial, and environmental sounds by image u-net. IEEE Access 11:144456–144483

Ha JS (2024) Video3D: single image to 3D using video diffusion and volume renderer. https://cs.brown.edu/media/filer_public/aa/4a/aa4a6cdf-6edf-4f1f-b628-7496e576504f/junsukha.pdf
Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, Yang Z, Zhang Y, Tao D (2020) A survey on vision transformer. IEEE Trans Pattern Anal Machine Intell 45:1
Hao X, Su X, Wang Z, Zhang H, Batushiren (2020) UnetGAN: a robust speech enhancement approach in time domain for extremely low signal-to-noise ratio condition. arXiv preprint arXiv:2010.15521
He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
Hodosh M, Young P, Hockenmaier J (2013) Framing image description as a ranking task: data, models and evaluation metrics. J Artif Intell Res 47:853–899
Hou Y, Ma X, Zhang J, Guo C (2024) Symmetric connected U-Net with multi-head self attention (MHSA) and WGAN for image inpainting. Symmetry 16(11):1423
Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861
Huang Y, Xu J, Jiang Z, Lai J, Li Z, Yao Y, Chen T, Yang L, Xin Z, Ma X (2023) Advancing transformer architecture in long-context large language models: a comprehensive survey. arXiv:2311.12351
Iandola FN, Moskewicz MW, Ashraf K, Han S, Dally WJ, Keutzer K (2016) Squeezenet: alexnet-level accuracy with 50x fewer parameters and <1mb model size. arXiv:1602.07360
Ibtehaz N, Rahman MS (2020) MultiresUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Netw 121:74–87
Iglesias G, Talavera E, Díaz-Álvarez A (2022) A survey on GANs for computer vision: recent research, analysis and taxonomy. Comput Sci Rev 48:100553 arXiv:abs/2203.11242
Ignacio MJ, Kim Y-G, Jin H, Yu S (2025) U-Net encapsulated transformer for reducing dimensionality in training large language models. https://github.com/ignaciomarvinjohn/uetlm. Accessed 2025
Ignacio MJ, Nguyen TT, Jin H, Kim Y-G (2024) Meme analysis using LLM-based contextual information and u-net encapsulated transformer. IEEE Access
Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. PMLR, pp 448–456
Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
Jansson A, Humphrey EJ, Montecchio N, Bittner R, Kumar A, Weyde T (2017) Singing voice separation with deep u-net convolutional networks. In: 18th international society for music information retrieval conference
Jiang Y, Tang H, Chang J-HR, Song L, Wang Z, Cao L (2023) Efficient-3dim: learning a generalizable single-image novel-view synthesizer in one day. arXiv:2310.03015
Karras JS, Holynski A, Wang T-C, Kemelmacher-Shlizerman I (2023) Dreampose: fashion image-to-video synthesis via stable diffusion. In: 2023 IEEE/CVF international conference on computer vision (ICCV), pp 22623–22633
Kaur A, Dong G, Basu A (2022) Gradxcepunet: explainable ai based medical image segmentation. In: International conference on smart multimedia. Springer, pp 174–188
Kerdvibulvech C (2025) Multimodal AI model for zero-shot vehicle brand identification. Multimedia Tools Appl 1–20
Khoee AG, Yu Y, Feldt R (2024) Domain generalization through meta-learning: a survey. Artif Intell Rev 57(10):285
Kim T-S, Ignacio MJ, Yu S, Jin H, Kim Y-G (2024) UI/UX for generative AI: taxonomy, trend, and challenge. IEEE Access 12:179891–179911
Lin T, Wang Y, Liu X, Qiu X (2021) A survey of transformers. AI Open 3:111–132
Liu H, Chen Z, Yuan Y, Mei X, Liu X, Mandic DP, Wang W, Plumbley MD (2023) Audioldm: text-to-audio generation with latent diffusion models. In: International conference on machine learning
Li D, Yang Y, Song Y-Z, Hospedales T (2018) Learning to generalize: meta-learning for domain generalization. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
Macartney C, Weyde T (2018) Improved speech enhancement with the wave-U-Net. arXiv preprint arXiv:1811.11307
Ma L, Jia X, Sun Q, Schiele B, Tuytelaars T, Gool LV (2017) Pose guided person image generation. arXiv:1705.09368
Majumder N, Hung C-Y, Ghosal D, Hsu W-N, Mihalcea R, Poria S (2024) Tango 2: aligning diffusion-based text-to-audio generations through direct preference optimization. In: ACM multimedia
Ma J, Li F, Wang B (2024) U-mamba: enhancing long-range dependency for biomedical image segmentation. arXiv:2401.04722
Na IS, Tran CD, Nguyen D, Dinh SV (2020) Facial UV map completion for pose-invariant face recognition: a novel adversarial approach based on coupled attention residual UNets. HCIS 10:1–17

Nagel M, Fournarakis M, Amjad RA, Bondarenko Y, Baalen M, Blankevoort T (2021) A white paper on neural network quantization. arXiv:2106.08295
Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 807–814
Nguyen HTT, Cao HQ, Nguyen KVT, Pham NDK (2021) Evaluation of explainable artificial intelligence: shap, lime, and cam. In: Proceedings of the FPT AI conference, pp 1–6
Omeiza D, Webb H, Jirotka M, Kunze L (2021) Explanations in autonomous driving: a survey. IEEE Trans Intell Transp Syst 23:10142–10162
Oord A, Kalchbrenner N, Espeholt L, Vinyals O, Graves A, Kavukcuoglu K (2016) Conditional image generation with pixelcnn decoders. Adv Neural Inf Process Syst 29
OpenAI (2023) ChatGPT (Mar 14 version). Large language model. https://chat.openai.com/chat
Poole B, Jain A, Barron JT, Mildenhall B (2022) Dreamfusion: text-to-3D using 2D diffusion. arXiv:2209.14988
Punn NS, Agarwal S (2021) Modality specific U-Net variants for biomedical image segmentation: a survey. Artif Intell Rev 55:5845–5889
Qin HX, Hui P (2023) Empowering the metaverse with generative AI: survey and future directions. In: 2023 IEEE 43rd international conference on distributed computing systems workshops (ICDCSW), pp 85–90
Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR, pp 8748–8763
Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. https://openai.com/blog/language-unsupervised/
Rahmon G, Bunyak F, Seetharaman G, Palaniappan K (2021) Motion U-Net: multi-cue encoder-decoder network for motion segmentation. In: 2020 25th international conference on pattern recognition (ICPR), pp 8125–8132
Ramachandran P, Zoph B, Le QV (2018) Searching for activation functions. arXiv:1710.05941
Raut G, Singh A (2024) Generative ai in vision: a survey on models, metrics and applications. arXiv:2402.16369
Raza R, Bajwa UI, Mehmood Y, Anwar MW, Jamal MH (2022) dResU-Net: 3D deep residual u-net based brain tumor segmentation from multimodal MRI. Biomed Signal Process Control 79:103861
Redko I, Morvant E, Habrard A, Sebban M, Bennani Y (2020) A survey on domain adaptation theory: learning bounds and theoretical guarantees. arXiv preprint arXiv:2004.11829
Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10684–10695
Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, Springer, pp 234–241
Ruan J, Xiang S (2024) VM-UNet: vision mamba UNet for medical image segmentation. arXiv:2402.02491
Ruiz N, Li Y, Jampani V, Pritch Y, Rubinstein M, Aberman K (2022) Dreambooth: fine tuning text-to-image diffusion models for subject-driven generation. In: 2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 22500–22510
Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, Ghasemipour K, Gontijo Lopes R, Karagol Ayan B, Salimans T et al (2022) Photorealistic text-to-image diffusion models with deep language understanding. Adv Neural Inf Process Syst 35:36479–36494
Sakai T, Nagai T (2021) Explainable autonomous robots: a survey and perspective. Adv Robot 36:219–238
Salimans T, Karpathy A, Chen X, Kingma DP (2017) Pixelcnn++: improving the pixelcnn with discretized logistic mixture likelihood and other modifications. arXiv preprint arXiv:1701.05517
Sharma N, Saba L, Khanna NN, Kalra MK, Fouda MM, Suri JS (2022) Segmentation-based classification deep learning model embedded with explainable AI for COVID-19 detection in chest X-ray scans. Diagnostics 12(9):2132
Shi Y, Tang H, Baine M, Hollingsworth MA, Du H, Zheng D, Zhang C, Yu H (2023) 3DGAUnet: 3D generative adversarial networks with a 3D U-Net based generator to achieve the accurate and effective synthesis of clinical tumor image data for pancreatic cancer. Cancers 15(23):5496
Shin S-Y, Kang Y-W, Kim Y-G (2020) Reward-driven U-Net training for obstacle avoidance drone. Expert Syst Appl 143:113064. https://doi.org/10.1016/j.eswa.2019.113064
Si C, Huang Z, Jiang Y, Liu Z (2023) Freeu: free lunch in diffusion u-net. In: 2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 4733–4743
Song CH, Han HJ, Avrithis Y (2021) All the attention you need: global-local, spatial-channel attention for image retrieval. In: 2022 IEEE/CVF winter conference on applications of computer vision (WACV), pp 439–448

Sun F, Li L, Qiu X, Liu Y (2018) U-Net: machine reading comprehension with unanswerable questions. arXiv preprint arXiv:1810.06638
tamasino52 (2021) UNETR-Pose. GitHub
Tang D, Chen Z (2024) English long sentence machine translation algorithm based on u-net generation adversarial network. In: 2024 second international conference on data science and information system (ICDSIS), pp 1–5
Tan M, Le QV (2019) EfficientNet: rethinking model scaling for convolutional neural networks. arXiv:1905.11946
Team G (2024) Gemini: a family of highly capable multimodal models. arXiv:2312.11805
Ulyanov D, Vedaldi A, Lempitsky V (2016) Instance normalization: the missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022
Umirzakova S, Ahmad S, Khan LU, Whangbo T (2024) Medical image super-resolution for smart healthcare applications: a comprehensive survey. Inf Fusion 103:102075
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA (2021) Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 79:102470
Wang J, Ignacio MJ, Yu S, Jin H, Kim Y-G (2024) UET4Rec: U-Net encapsulated transformer for sequential recommender. Expert Syst Appl 255:124781
Wang Z, Zheng J-Q, Zhang Y, Cui G, Li L (2024) Mamba-UNet: UNet-like pure visual mamba for medical image segmentation. arXiv:2402.05079
Wu D-Y, Chen Y-H, Lee H-Y (2020) Vqvc+: one-shot voice conversion by vector quantization and u-net architecture. arXiv preprint arXiv:2006.04154
Wu Y, He K (2018) Group normalization. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
Wu J, Liu W, Li C, Jiang T, Shariful IM, Sun H, Li X, Li X, Huang X, Grzegorzek M (2022) A state-of-the-art survey of u-net in microscopic image analysis: from simple usage to structure mortification. arXiv:2202.06465
Wu L, Xia Y, Zhao L, Tian F, Qin T, Lai J, Liu T-Y (2017) Adversarial neural machine translation. arXiv:abs/1704.06933
Xie S, Zhang Z, Lin Z, Hinz T, Zhang K (2022) Smartbrush: text and shape guided object inpainting with diffusion model. In: 2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 22428–22437
Xing Z, Ye T, Yang Y, Liu G, Zhu L (2024) Segmamba: long-range sequential modeling mamba for 3D medical image segmentation. In: International conference on medical image computing and computer-assisted intervention
Xu B, Wang N, Chen T, Li M (2015) Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853
Yang Z, Chen W, Wang F, Xu B (2017) Improving neural machine translation with conditional sequence generative adversarial nets. arXiv:1703.04887
Yan Z, Li X, Li M, Zuo W, Shan S (2018) Shift-net: image inpainting via deep feature rearrangement. In: European conference on computer vision
Yeong DJ, Panduru K, Walsh J (2025) Exploring the unseen: a survey of multi-sensor fusion and the role of explainable AI (XAI) in autonomous vehicles. Sensors 25(3):856
Yin Y, Xu D, Wang X, Zhang L (2021) AGUNet: annotation-guided U-Net for fast one-shot video object segmentation. Pattern Recognit 110:107580
Zeng F, Gan W, Wang Y, Yu PS (2023) Distributed training of large language models. In: 2023 IEEE 29th international conference on parallel and distributed systems (ICPADS), pp 840–847
Zhang Z, Liu Q, Wang Y (2018) Road extraction by deep residual u-net. IEEE Geosci Remote Sens Lett 15(5):749–753
Zhang L, Yang H, Qiu T, Li L (2022) AP-GAN: improving attribute preservation in video face swapping. IEEE Trans Circuits Syst Video Technol 32:2226–2237
Zhang L, Rao A, Agrawala M (2023) Adding conditional control to text-to-image diffusion models. In: 2023 IEEE/CVF international conference on computer vision (ICCV), pp 3813–3824
Zhang X, Zhou X, Lin M, Sun J (2017) Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 6848–6856
Zhou P, Wang L, Liu Z, Hao Y, Hui P, Tarkoma S, Kangasharju J (2024) A survey on generative ai and LLM for video generation, understanding, and streaming. arXiv:2404.16038
Zhou D, Wang W, Yan H, Lv W, Zhu Y, Feng J (2022) Magicvideo: efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018

Zhou S, Yang P, Wang J, Luo Y, Loy CC (2024) Upscale-a-video: temporal-consistent diffusion model for real-world video super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2535–2545

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Marvin John Ignacio[1] · Sangyun Shin[2] · Hulin Jin[3] · Seong Joon Yoo[1] · Dongil Han[1] · Yong-Guk Kim[1]

✉ Yong-Guk Kim
ykim@sejong.ac.kr

Marvin John Ignacio
mjci@sju.ac.kr

Sangyun Shin
sangyun.shin@cs.ox.ac.uk

Hulin Jin
jinhulin@ahu.edu.cn

Seong Joon Yoo
sjyoo@sejong.ac.kr

Dongil Han
dihan@sejong.ac.kr

[1] Department of Computer Engineering, Sejong University, Gwangjin-Gu, Seoul 05006, Republic of Korea

[2] Department of Computer Science, University of Oxford, 7 Parks Road, Oxford OX1 3QG, UK

[3] School of Computer Science and Technology, Anhui University, Hefei, China