DiCo-Net: A Two-Stage Diffusion-Convolution Network with Multimodal Fusion for High-Resolution line-art Extraction

Wanchuang Luo[†]
Northwest A&F University
Yangling, Shaanxi, China
luowanchuang@nwafu.edu.cn

Fangbo Lu[†]
Northwest A&F University
Yangling, Shaanxi, China
fangbolu@nwafu.edu.cn

Yongyuan Qiao
Northwest A&F University
Yangling, Shaanxi, China
qiaoyongyuan@nwafu.edu.cn

Meili Wang[*]
Northwest A&F University
Yangling, Shaanxi, China
wml@nwsuaf.edu.cn

Abstract

Line-art extraction lies at the intersection of computer vision and digital art, seeking to recover semantically coherent contours from natural images to support downstream applications such as anime colorization and style transfer. Classical edge detectors are notoriously sensitive to noise and often fail to separate semantic boundaries from textural artifacts, while CNNs remain vulnerable to illumination and stylistic variations. Diffusion-based approaches, although promising, frequently yield unstable outputs that do not meet artistic requirements. To address these limitations, we propose DiCo-Net, a two-stage hybrid framework that integrates diffusion models with CNN-based refinement. In the first stage, a pretrained diffusion model (the base model) generates a preliminary line drawing. The second stage employs an encoder–decoder CNN (the refine model) to enhance structural fidelity and suppress residual artifacts. A multimodal fusion module injects auxiliary cues—specifically, texture maps and depth maps extracted from the input image—into the network, thereby strengthening semantic disambiguation. In addition, Linear–Kan and non-uniform sampling convolutional kernels are introduced to improve multi-scale feature modeling. To evaluate the proposed method, we curate LINE-2K, a high-resolution photo/line-art paired dataset comprising 4,000 hand-drawn pairs; after data augmentation, the corpus expands to 20,000 pairs and is partitioned into training, validation, and test sets with a 6:2:2 split. Extensive experiments on LINE-2K demonstrate that DiCo-Net preserves semantic contours while rendering rich textures and high artistic quality, consistently outperforming representative baselines.

Keywords: Line-art Extraction, Diffusion Model, Convolutional Network, Multimodal fusion

1. Introduction

Line-art extraction is an important problem in computer vision that seeks to recover semantically coherent, structurally faithful, and aesthetically pleasing contour maps from raw images. Beyond supporting image feature modeling and structural understanding, high-quality line drawings serve as crucial inputs for tasks such as animation colorization, style transfer, and digital illustration. The central challenge is to balance visual aesthetics and texture expressiveness while preserving structural integrity and fine detail.Classical edge detectors laid the early technical groundwork for line-art extraction. Traditional operators [18, 14, 1] leverage grayscale gradients and second-order derivatives to detect edges at low-level pixel scales. However, their reliance on fixed thresholds and filters renders them highly sensitive to noise and prone to confusing semantic boundaries with local textural artifacts. In complex scenes, traditional operators often produce fragmented strokes or over-responses, falling short of artistic line-art requirements.

The advent of deep learning has substantially advanced edge detection performance. Convolutional Neural Network methods [23, 12, 21] leverage multi-scale feature fusion and end-to-end training, with lightweight architectures and multi-scale modeling strategies further optimizing efficiency and detail rendering. Recently, Transformer-based approaches [16, 5, 28] utilize global self-attention to capture long-range dependencies, improving global structural modeling. Nevertheless, they incur substantial computational overhead and still underutilize texture and depth

[†]These authors contributed equally.
[*]Corresponding author. Email: wml@nwsuaf.edu.cn

cues in artistic settings. Diffusion models [3, 26, 30, 7], benefiting from their progressive denoising generation process, demonstrate advantages in pixel-level detail expression and global consistency. However, they suffer from high computational cost, slow inference speeds unsuitable for interactive use, and instability in complex artistic images—manifesting as over-smoothing or detail loss.

To address these challenges, we propose DiCo-Net, a two-stage hybrid line-art extraction framework that integrates diffusion models with convolutional refinement. In stage one, a pretrained diffusion model (Kontext [7]) generates a preliminary line drawing, leveraging strong global consistency and detail modeling to establish the foundational contours. In stage two, an encoder–decoder CNN refines the result. A multimodal fusion module incorporates texture and depth maps [25] derived from the original image as auxiliary inputs, enhancing semantic continuity and textural expressiveness. To further improve multi-scale capacity, we introduce non-uniformly sampled convolutional kernels inspired by SKConv and combine them with linear-kan techniques to flexibly model features across scales. We validate the approach on LINE-2K, a high-resolution photo/line-art paired dataset curated for this study. Experimental results show that our method preserves clear semantic contours while producing rich textures and artistic style, significantly outperforming existing baselines.

2. Related work

Existing line-art extraction methods can be broadly categorized into four classes: classical edge detectors, convolutional neural network methods, Transformer-based approaches, and diffusion models. Classical edge detectors laid the early technical groundwork for line-art extraction. Operators such as Sobel[18], Canny[1], and LoG (Laplacian of Gaussian)[14] leverage grayscale gradients and second-order derivatives to detect edges at low-level pixel scales. However, their reliance on fixed thresholds and filters renders them highly sensitive to noise and prone to confusing semantic boundaries with local textural artifacts. In complex scenes, traditional operators often produce fragmented strokes or over-responses, falling short of artistic line-art requirements.

Deep convolutional neural networks have substantially advanced edge detection and line-art extraction. Xie *et al*. [23] proposed HED (Holistically-Nested Edge Detection), an end-to-end edge detection framework that fuses multi-scale features to produce cleaner edge maps, while Liu *et al*. [12] proposed RCF, an improved edge detection framework that strengthens performance through richer convolutional representations and multi-scale feature aggregation. Su *et al*. [21] developed PiDiNet (Pixel Difference Network), combining the interpretability of traditional edge operators with the representation power of CNNs through pixel-difference convolutions, despite its lightweight design, it surpasses human-perceived edges on BSDS500. Efficiency-oriented models—including DexiNed (Dense Extreme Inception Network) [15], LDC (Lightweight Dense CNN) [20], and TEED (Tiny and Efficient Edge Detector) [19]—propose ultra-compact architectures (fewer than one million parameters) that approach the accuracy of larger models. Other efforts emphasize multi-scale modeling and hierarchical fusion. He *et al*. [4] introduced BDCN (Bi-Directional Cascade Network), where layer-wise supervision guides each layer to capture scale-specific structures, and the proposed scale enhancement module further boosts multi-scale edge representation, yielding state-of-the-art performance. Xuan *et al*. [24] proposed FCL-Net, which introduces a fine-scale corrective learning mechanism where deep semantic features guide shallow layers, thereby enhancing the capture of fine-grained edge details. To address subjectivity and class imbalance in sketch annotations, Cetinkaya *et al*. [2] introduced RankED, using a ranking-based loss to handle both imbalance and uncertainty, improving edge distinction in challenging areas. Zhou *et al*. [29] developed UAED, leveraging learnable Gaussian label distributions combined with uncertainty-weighted loss functions to better handle hard or ambiguous edge samples, while Li *et al*. [8] developed Beta Network, leveraging the Beta distribution to explicitly capture boundary uncertainty in edge detection. More recently, Li *et al*. [10] proposed the Doubly Decoupled Network, which enhances CNN-based edge detection by employing decoupling strategies at both the data and feature levels: at the data level, learnable Gaussian sharpening strengthens edge evidence; at the feature level, shallow features are decomposed into spatial and semantic components to reduce redundancy, yielding leading results across several benchmarks. Despite these advances, the locality of convolutional kernels limits global structure modeling, making CNNs sensitive to illumination and style variations and prone to trading off semantic contours against textural detail—often causing contour loss or excessive noise retention in stylized imagery.

Transformers and self-attention have recently been explored for edge detection. Unlike CNNs, Transformers[22] capture long-range dependencies through global self-attention, which is advantageous for modeling global semantic structure. EDTER as proposed by Pu *et al*. [16] employs a two-stage design in which a global Transformer encoder captures long-range context and a local Transformer refines fine-grained regions; combined with bidirectional multi-layer aggregation, it delivers clear edges with global consistency. Huan *et al*. [5] introduced CATS (Context-Aware Tracing Strategy), which mitigates feature co-occurrence common in CNNs via feature de-mixing and context-aware fusion, improving localization. MuGE (Mul-

tiple Granularity Edge Detection) [28] addresses annotation subjectivity by generating edge maps at multiple granularities—from coarse outlines to fine textures—adapting well to artistic line-art scenarios. EdgeNAT [6] and SAUGE [13] further improve performance and controllability via efficient Transformer designs and multi-granularity fusion. Nevertheless, Transformer-based methods typically entail high computational cost and still underutilize texture and depth cues in artistic settings.

Diffusion models, which have achieved notable success in image generation and editing, have also been applied to edge and sketch extraction [3]. Unlike CNNs or Transformers, diffusion methods synthesize targets via progressive denoising, enabling detail-rich, pixel-level outputs. DiffusionEdge as proposed by Ye *et al*. [26], which pioneers diffusion probability models for general edge detection, predicting edges at the original image scale to avoid encoder–decoder blurring and achieving superior accuracy on NYUDv2. Generative Edge Detection (GED) [30], based on Stable Diffusion, further demonstrates the potential of diffusion for high-quality line-art generation. FLUX.1 Kontext [7] integrates semantic context in latent space to enforce local and global consistency, highlighting strengths in both pixel-level detail and global coherence. However, diffusion approaches often incur high computational cost, slow inference unsuitable for interactive use, and instability in artistic scenarios—manifesting as over-smoothing or detail loss.

DiCo-Net Network Architecture As shown in Fig. 1, the overall architecture of the DiCo-Net network consists of two major modules: the Base Model and the Refine Model, forming a two-stage processing workflow from preliminary sketch generation to fine-grained refinement.

3. Proposed network

3.1. Overall network architecture

The core function of the Base Model is to generate preliminary line sketches (base sketches). It employs a pretrained diffusion network to handle pixel fluctuations in photos caused by varying external conditions. The base sketch, along with texture maps and depth maps, is input into the Multimodal Line Enhancement Module (MEM). The MEM module fuses structural information, texture features, and depth information, providing richer input representations for the subsequent encoder.

The network employs Residual U-block (RSU) [17] for feature extraction. RSU captures features across different receptive fields through residual connections, effectively merging local details with global structural information. This addresses the issue of edge information loss in fine-grained lines commonly encountered by traditional convolutional networks. Additionally, an Uneven Selective

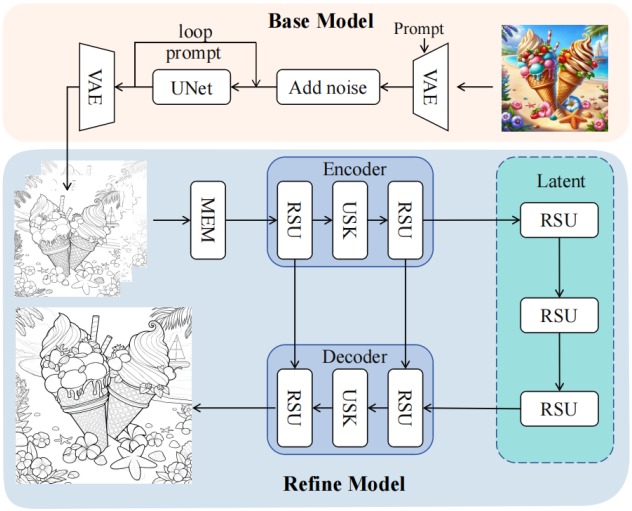

Figure 1. Network architecture

Kernel Convolution (USK) [9] module is integrated within the encoder. USK employs a dynamic selection strategy for non-uniform multi-scale convolution kernels, adaptively capturing line features across different scales. This enables the network to preserve detail clarity while maintaining line continuity and overall structural coherence.

The extracted multi-scale features are then mapped to a latent space for semantic exchange, achieving global information integration. This process guides local details within a global context, preventing isolated lines, false responses, and structural discontinuities. Finally, the decoder remaps the semantically exchanged latent features back to the pixel space, generating a refined line-art image.

The core challenge in high-quality line-art extraction lies in preserving semantic contour continuity while restoring fine-grained textures and artistic style. Traditional methods face limitations when processing hand-drawn or anime images: noise interference easily confuses semantic edges with non-critical textures; low-quality line-art often suffers from texture loss, struggling to represent perspective-based line thickness variations and diminishing artistic expressiveness.

To address this, we propose the MEM module, which enhances line-art expressiveness through multimodal feature fusion and context-aware convolutions. MEM's texture embedding supplements fine-grained texture details, while its depth embedding captures artistic style features, enabling line thickness to vary with distance. Drawing inspiration from the Contextual Transformer (CoT) [11] architecture, MEM employs self-attention mechanisms to jointly represent texture and depth information, improving overall structural integrity and detail rendering.

line-art incorporates subtle brushstrokes and varying contour thicknesses, requiring models to be sensitive to lo-

cal features while maintaining global coherence. To address this, the UCA-SKConv (USK) module was designed, employing non-uniform sampling convolutional kernels and self-attention mechanisms for multi-scale feature fusion. USK employs convolutional kernels of varying sizes (1, 3, 7) to extract multi-level features. Attention mechanisms highlight critical line regions while suppressing background noise, yielding clearer results.

Additionally, USK incorporates the Linear-KAN structure during channel weight generation. This uses learnable B-spline basis functions to enhance channel-level nonlinear representation capabilities, enabling more precise attention weight generation.

To fully leverage the network's multi-layer feature representation capabilities, deep supervision is introduced on both lateral outputs, optimized using binary cross-entropy loss. The low-order lateral output captures local details, while the high-order lateral output focuses on semantic structure. The total loss is obtained through weighted summation to balance consistency between detail and structure.

3.2. MEM

The central challenge in high-quality line-art extraction is to reconstruct fine-grained textures while maintaining the continuity of semantic contours and preserving artistic style. Conventional methods exhibit notable shortcomings on hand-drawn and anime imagery: (i) noise readily confounds semantic boundaries with nonessential textures, and (ii) low-quality outputs often lose texture detail and violate perspective-driven line-thickness cues—producing overly thin foreground strokes and excessively thick background strokes—thereby diminishing artistic expressiveness.

To address the aforementioned issues, this study introduces the MEM module, which enhances the representational capacity of line art through multimodal feature fusion and context-aware convolutional modeling. As shown in Fig. 2, the texture embedding branch of MEM integrates a texture map $T(X)$, computed from image gradients, into the feature space. This alleviates texture loss during the base model's line-art generation process, thereby restoring fine-grained high-frequency details. Leveraging the implicit feature filtering capability of convolutional neural networks, this work employs the classical Canny edge detector to extract rich, albeit partially redundant, high-frequency texture features. Subsequent convolutional self-attention enables adaptive discrimination and effective fusion of texture-related and semantic information. Furthermore, recognizing that stylized line art typically exhibits distinct foreground–background hierarchies, MEM incorporates a depth embedding branch to provide explicit depth priors. This branch guides line-art generation toward perspective-consistent stroke rendering, producing thicker and more prominent strokes in foreground regions and finer,

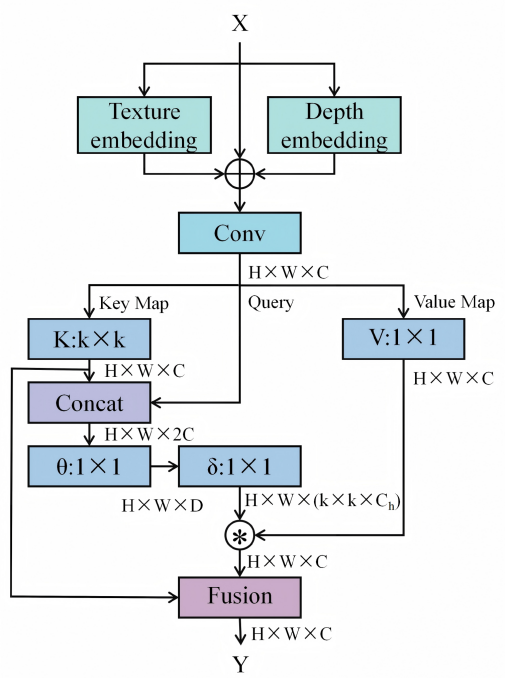

Figure 2. Structure of the MEM module.

softer lines in distant or background areas. Depth maps are batch-generated offline using a pretrained Depth Anything model and cached to balance representational richness with training efficiency. By combining static and dynamic context modeling based on a Contextual Transformer (CoT) mechanism and jointly modeling texture and depth features via self-attention, MEM enhances structural coherence while preserving fine-grained detail representation at the feature level.

In implementation, MEM first performs multimodal feature embedding on the input image $X \in \mathbb{R}^{H\times W\times C}$:

$$F_{\text{emb}} = \text{Concat}(X, T(X), D(X)) \in \mathbb{R}^{H\times W\times C} \quad (1)$$

where $T(X)$ and $D(X)$ denote the texture and depth maps, respectively. The embedded features are then processed by convolutional layers to extract preliminary representations that capture low-level texture and structural information:

$$F_0 = F_{\text{conv}}(F_{\text{emb}}) \quad (2)$$

Subsequently, MEM enhances feature representation within both static and dynamic contexts by referencing the CoT attention mechanism. First, static contextual features are extracted using a 3×3 convolution:

$$K_{\text{static}} = \text{Conv}_{3\times 3}(F_0) \quad (3)$$

which captures local structural patterns and provides a foundational representation for dynamic attention. The preliminary feature F_0 is then concatenated with the static context K_{static} and passed through two successive 1×1 convolution layers to generate the attention matrix $\mathcal{A}$, which is

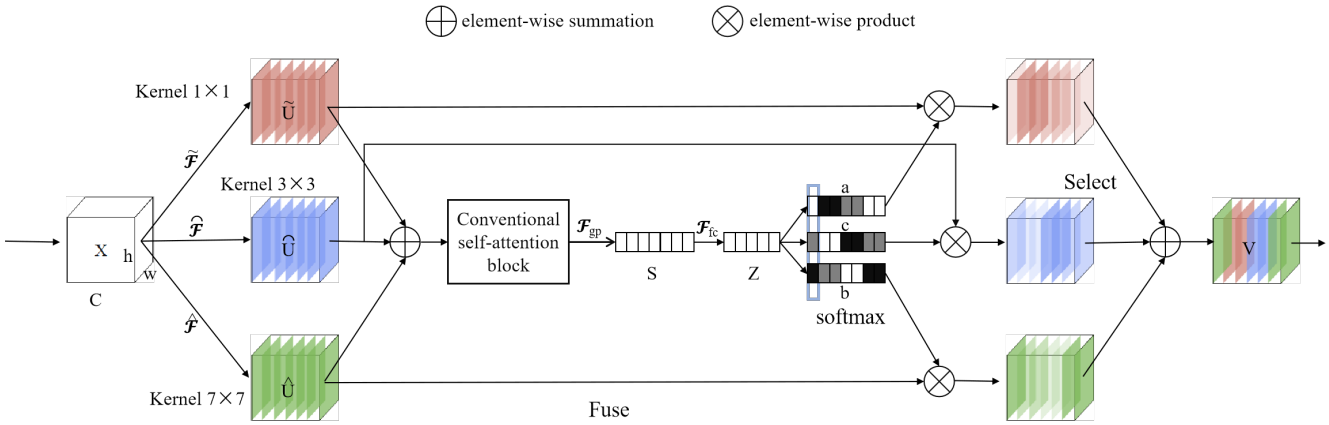

Figure 3. USK framework.

multiplied by the value mapping V to obtain the dynamic context:

$$K_{\text{dynamic}} = \text{softmax}(\mathcal{A}) \odot V \tag{4}$$

$$\mathcal{A} = \text{Conv}_{1\times1}([F_0, K_{\text{static}}]) \tag{5}$$

$$V = \text{Conv}_{1\times1}(F_0) \tag{6}$$

Here, $\mathcal{A}$ serves to weight the relationships among multimodal features, while V provides adaptive value features, effectively capturing the interactions between texture and depth information. Finally, the static and dynamic contextual features are combined to form the enhanced representation F_{att}, which is fused with the preliminary feature F_0 to produce the output Y:

$$F_{\text{att}} = K_{\text{static}} + K_{\text{dynamic}} \tag{7}$$

$$Y = F_{\text{fusion}}(F_0, F_{\text{att}}) \in \mathbb{R}^{H\times W\times C} \tag{8}$$

3.3. USK

Line-art comprises extremely fine strokes and contours with varying thickness, requiring models to be highly sensitive to local features while preserving global coherence. To meet this requirement, we propose the UCA-SKConv module (abbreviated USK), which couples non-uniformly sampled convolutional kernels with self-attention, drawing inspiration from Selective Kernel Convolution (SKConv). As illustrated in Fig. 3, USK integrates multi-scale modeling with dynamic attention selection, enabling robust capture of stroke thickness under noise while maintaining holistic structural consistency.

The Split stage in USK adopts non-uniform kernel sizing $\{1, 3, 7\}$ to match the heterogeneous statistics of line-art strokes. Small kernels 1×1 emphasize ultra-fine strokes, mitigating smoothing and detail loss during convolution; medium kernels 3×3 extract mid-scale structural cues and serve as the most versatile branch for line-art extraction; large kernels 7×7 capture broad context, promoting continuity of thick strokes and global contours. This non-uniform design reduces redundant computation relative to uniform scale partitioning and improves precise responses to lines at different scales, thereby balancing detail preservation and structural integrity.Let the input feature be denoted as $\mathbf{X_1} \in \mathbb{R}^{B\times C\times H\times W}$. The output of each branch in the Split module can be expressed as

$$\mathbf{F} * i = \text{Conv} * k\_i(\mathbf{X\_1}) \quad k\_i \in 1, 3, 7 \tag{9}$$

This non-uniform design effectively mitigates the redundant computations commonly caused by uniform scale partitioning and enhances the model's ability to precisely respond to line structures at varying levels. Consequently, it achieves a balanced trade-off between detail preservation and structural consistency.

Conventional SKConv relies primarily on channel-wise weighting and thus underutilizes spatial positional information, which is critical for line-art extraction. To address this limitation, we insert a self-attention mechanism at feature concatenation,

$$\mathbf{F}_U = \text{Concat}\left(\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_i\right) \tag{10}$$

enabling the model to adaptively emphasize salient regions based on spatial dependencies. Specifically, $\mathbf{F}_U$ is projected into query, key, and value matrices $\mathbf{Q}, \mathbf{K}, \mathbf{V}$, and the attention matrix is computed as

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right) \tag{11}$$

yielding the enhanced feature representation $\mathbf{F}_{\text{attn}} = \mathbf{AV}$. By explicitly modeling spatial dependencies, this mechanism effectively highlights key line regions while suppressing background noise, resulting in clearer extracted contours and cleaner backgrounds.

In the channel weighting stage, conventional SKConv employs a multi-layer perceptron (MLP) to process globally pooled features. However, its linear mapping scheme

provides limited capability for modeling local patterns. To better support the adaptive modeling requirements of non-uniform convolution kernels, the MLP in the SK module is replaced with Linear-KAN. This design choice was empirically observed to yield stable performance improvements in preliminary experiments and is therefore adopted in the final model. Specifically, the proposed USKConv replaces the MLP with a Linear-KAN module, which expands each globally pooled scalar using learnable B-spline basis functions and then linearly combines them to generate branch weights. First, global average pooling is performed along the channel dimension:

$$\mathbf{s} = \mathrm{GAP}(\mathbf{F}_{\text{attn}}) \in \mathbb{R}^C \tag{12}$$

For each channel s_c, K learnable B-spline bases $\{B_k(s_c)\}_{k=1}^K$ are introduced to construct a basis expansion, followed by a linear combination to compute the branch scores:

$$z_i = \sum_{c=1}^{C} \sum_{k=1}^{K} W_{i,c,k}\, B_k(s_c) + b_i \quad i = 1, \ldots, M \tag{13}$$

which are then normalized via a softmax function to obtain the attention weights:

$$\mathbf{a}_i = \frac{\exp(z_i)}{\sum_{j=1}^{M} \exp(z_j)} \tag{14}$$

Finally, the branch features are fused through weighted summation:

$$\mathbf{F}_V = \sum_{i=1}^{M} \mathbf{a}_i\, \mathbf{F}_i \tag{15}$$

Compared with traditional MLP-based implementations, Linear-KAN introduces only marginal computational overhead while incorporating nonlinear channel mappings through B-spline basis functions. This design substantially enhances the local pattern modeling capability in attention weight generation, enabling the network to more accurately select branch features that are most beneficial for precise line-art extraction.

3.4. Loss Function

To fully exploit the network's hierarchical feature representations, we impose deep supervision on two side outputs and optimize them using binary cross-entropy (BCE). The lower-order side output emphasizes local details and fine contours, whereas the higher-order side output aggregates semantic context and enforces structural coherence. Supervising predictions at multiple depths promotes global line continuity and structural integrity while preserving pixel-level accuracy.

Let the side-output prediction be denoted by S_{side} and the corresponding ground-truth line-art by Y. The BCE loss for a side output is defined as

$$\mathcal{L}_{\text{side}} = -\frac{1}{N} \sum_{i=1}^{N} \Big[y_i \log \hat{S}_{\text{side},i} + (1 - y_i) \log\big(1 - \hat{S}_{\text{side},i}\big) \Big] \tag{16}$$

where $\hat{S}_{\text{side}}$ denotes the predicted probabilities, $y_i \in \{0, 1\}$ is the binary label for pixel i, and N is the total number of pixels.

The overall training objective is a weighted sum of the losses from the two side outputs:

$$\mathcal{L}_{\text{total}} = \alpha\, \mathcal{L}_{\text{side}}^{\text{low}} + \beta\, \mathcal{L}_{\text{side}}^{\text{high}} \tag{17}$$

where α and β control the relative contributions of the lower- and higher-order side outputs, respectively, thereby balancing fine-grained detail with high-level semantic consistency.

4. Experiments and Results

4.1. Model configuration

We first describe the configuration of the two-stage model. The base model adopts the FLUX.1 Kontext image editing model, a 12 B parameter diffusion backbone. Compared with the 20 B parameter Qwen-Image-Edit-2509 and Qwen-Image-Edit-2511 models, FLUX.1 Kontext was selected due to its superior structural consistency in image-to-image transformations and lower inference latency. It is worth noting that the base model operates in an inference-only setting, the diffusion backbone is fully frozen and used for offline batch inference, requiring no additional training or fine-tuning. In preliminary experiments, we evaluated several image editing models, among which FLUX.1 Kontext exhibited the most favorable initial performance for line-art extraction. During inference, a fixed prompt ("Convert the image to a line drawing while keeping the structure and details unchanged") is employed, the number of inference steps is set to 25. Compared with partially fine-tuned or fully trained diffusion models, this configuration significantly reduces computational and training costs. The refinement model contains 21.6 M parameters. However, due to multi-layer feature caching, its computational complexity reaches 80.3 GFLOPS, which is comparable to that of existing line-art extraction models.

4.2. Datasets and Experimental Setup

To assess the effectiveness of the proposed approach for high-quality line-art extraction, we curate a high-resolution photo–line-art paired dataset, LINE-2K. The dataset comprises 4000 pairs of hand-drawn line-art and corresponding source images, which are expanded via data augmentation

to 20000 pairs to improve model generalization. We partition the data into training, validation, and test sets with a 6:2:2 ratio, ensuring that the model sufficiently learns line-art characteristics during training while enabling objective evaluation during validation and testing.

All experiments are conducted on a system equipped with an NVIDIA GeForce RTX 4060 Ti GPU and CUDA 12.9. Unless otherwise specified, the training hyperparameters are as follows: batch size of 5 for training and 1 for validation, an initial learning rate of 0.001, and the Adam optimizer for parameter updates.

4.3. Evaluation Metrics

We adopt a comprehensive, multi-dimensional evaluation protocol to quantify performance in terms of edge accuracy, structural consistency, and perceptual quality. Specifically, we report:

Pixel-wise accuracy. Mean Absolute Error (MAE) and Mean Squared Error (MSE) are computed after normalizing pixel intensities to $[0, 1]$, thereby measuring the discrepancy between generated line-art and ground-truth (GT) annotations at the pixel level and reflecting the fidelity of edge strength and contour shape.

Threshold-based detection quality. Optimal Image Scale (OIS) and Optimal Dataset Scale (ODS) summarize F-score performance under varying binarization thresholds. OIS captures per-image optimal performance (local optimum), whereas ODS evaluates a single, dataset-wide threshold (global consistency), thereby characterizing both instance-level and holistic detection behavior.

Perceptual quality. Because traditional numerical measures may not fully capture visual realism in line-art extraction, we additionally report the Learned Perceptual Image Patch Similarity (LPIPS) [27] , which correlates with human judgments of continuity, structural integrity, and naturalness.

This three-pronged protocol jointly addresses pixel-level precision, structural correspondence, and perceptual quality, providing an interpretable and practically meaningful assessment of overall model performance for line-art extraction.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{18}$$

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{19}$$

4.4. Comparative test and result analysis

Fig. 4 present a systematic comparison of Canny, TEED, UAED, DiffusionEdge, MuGE, and DDN on line-art extraction using five image sets from the LINE-2K dataset. All methods accomplish basic line detection; however, they differ markedly in detail preservation and visual fidelity. The classical Canny operator exhibits weak performance in fine-detail recovery and line continuity. The deep model TEED improves local texture retention via multi-scale feature fusion, yet its outputs still contain fractures and discontinuities in primary structures, reflecting limited global context modeling. UAED achieves robust global structure perception but insufficiently suppresses background textures, yielding spurious edges. DiffusionEdge strengthens global consistency through diffusion-based modeling but performs poorly in thin-line regions, causing local blurring and detail loss. MuGE captures global layouts and local details through multi-granularity fusion, but its edge completeness and perceptual quality remain constrained. DDN reduces detail fragmentation by decoupling trunk and detail edges and applying dense feature fusion; nevertheless, its results still show coarse textures and nonuniform stroke widths. In contrast, the proposed method provides stronger edge preservation, more effective suppression of background artifacts, and higher overall line quality, producing outputs that closely resemble professionally authored line-art in structural integrity and aesthetic coherence.

As summarized in Table 1, the proposed method surpasses existing approaches on LINE-2K. For pixel-level reconstruction accuracy, it attains the lowest MSE (0.0401) and MAE (0.0817), indicating higher numerical consistency in restoring edge intensity and contour geometry. For structural matching, it achieves the highest OIS (0.9610) and ODS (0.9602) among all baselines, demonstrating superior modeling of edge continuity and global topology under multi-threshold evaluation. For perceptual quality, its LPIPS score (0.2001) is substantially lower than those of competing methods (e.g., DDN: 0.3712; MuGE: 0.3785), implying closer perceptual similarity to the ground truth. Overall, the method advances the state of the art across pixel-level accuracy, structural consistency, and perceptual fidelity.

4.5. Ablation experiment

To assess the effectiveness of the proposed components, we conduct five ablation settings based on DiCo-Net, a two-stage hybrid line-art extraction model that integrates diffusion models with convolutional networks. The output of the first (diffusion) stage—denoted as Context—serves as the Base Model. Adding the original convolutional network yields the two-stage Baseline. We then augment the Baseline with the multimodal fusion module (MEM) and the adaptive convolution module (USK) separately to quantify their independent contributions. Finally, we incorporate both modules to obtain the full DiCo-Net and evaluate the aggregate performance gain. The qualitative results are summarized in Fig. 5, and the corresponding quantitative statistics are reported in Table 2. The observations can be summarized as follows.

Table 1. Performance comparison of different models on the LINE-2K dataset.

Model	MSE↓	MAE↓	OIS↑	ODS↑	LPIPS↓
Canny	0.0592	0.0886	0.9559	0.9545	0.4453
TEED	0.0977	0.1566	0.9550	0.9545	0.4907
UAED	0.0721	0.1380	0.9582	0.9575	0.4589
DiffusionEdge	0.0537	0.0856	0.9578	0.9572	0.3844
MuGE	0.0475	0.0949	0.9592	0.9586	0.3785
DDN	0.0447	0.0830	0.9602	0.9596	0.3712
Ours	**0.0401**	**0.0817**	**0.9610**	**0.9602**	**0.2001**

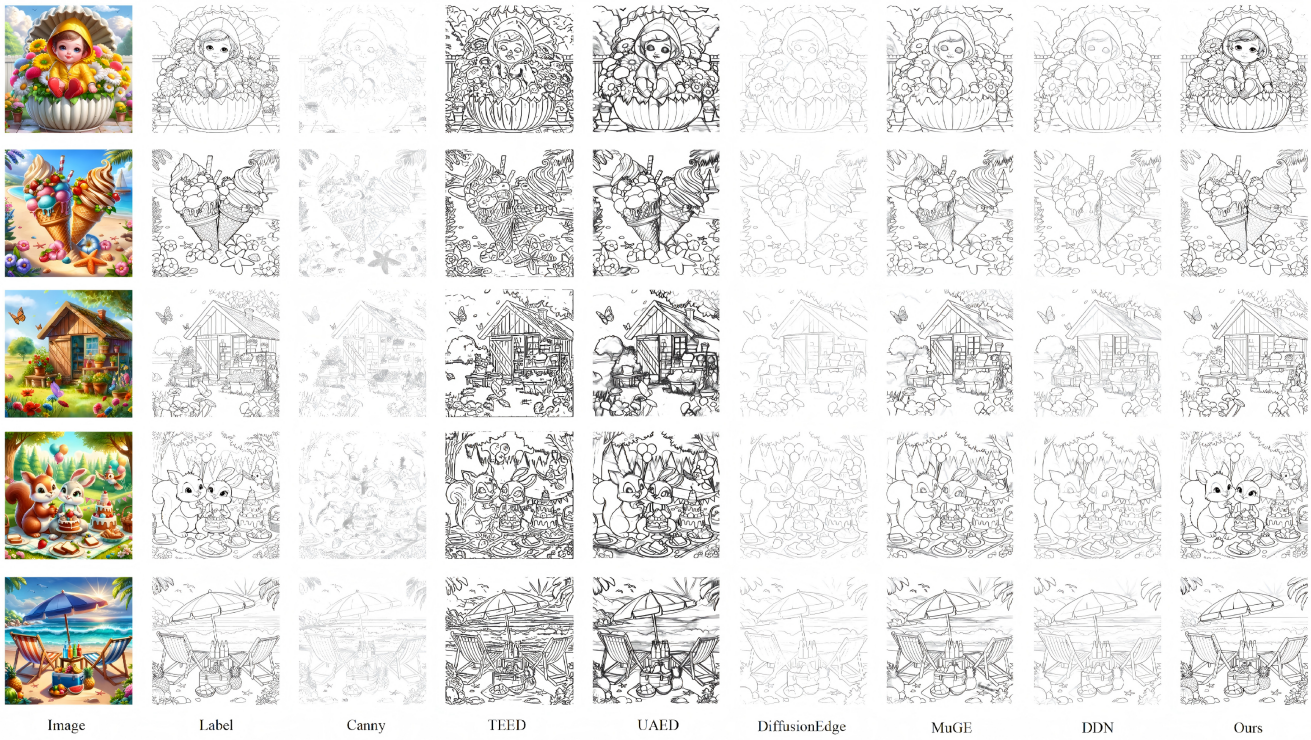

Figure 4. Comparison of Inference Results Across Different Models

The observations are as follows.Subsequently (i) The Base Model relies solely on a single-stage diffusion process, achieving MSE, MAE, OIS, ODS, and LPIPS scores of 0.0456, 0.0843, 0.9594, 0.9589, and 0.3535, respectively. These quantitative results suggest that, although the model is capable of producing coarse structural outlines, it suffers from notable limitations in pixel-wise reconstruction fidelity, structural coherence, and perceptual fidelity. (ii) The two-stage baseline model incorporates a convolutional refinement network, leading to consistent performance improvements. Specifically, MSE and MAE are reduced to 0.0441 and 0.0837, respectively, while OIS and ODS increase to 0.9597 and 0.9591, and LPIPS decreases to 0.3445. These results indicate that the refinement stage effectively mitigates noise artifacts introduced during the diffusion process and improves local edge fidelity. (iii) With the incorporation of MEM, the model attains MSE, MAE, OIS, ODS, and LPIPS scores of 0.0423, 0.0831, 0.9601, 0.9594, and 0.2724, respectively. These results demonstrate that MEM effectively enhances perceptual fidelity and semantic plausibility by integrating multimodal semantic cues with deep feature representations, while also yielding consistent improvements in pixel-wise reconstruction error and structural metrics. (iv) With the incorporation of the USK module, the model attains MSE, MAE, OIS, ODS, and LPIPS scores of 0.0417, 0.0825, 0.9607, 0.9598, and 0.2463, respectively. These results demonstrate that adaptive non-uniform convolutional kernels effectively enhance line continuity and geometric stability, thereby further improving the overall quality of the generated line art. (v) The proposed Base + Refine framework achieves the best performance across all evaluation metrics when all modules are jointly integrated. Specifically, it attains MSE, MAE, OIS, ODS, and LPIPS scores of 0.0401, 0.0817, 0.9610,

Table 2. Ablation experiment

Model	MSE↓	MAE↓	OIS↑	ODS↑	LPIPS↓
Base Model	0.0456	0.0843	0.9594	0.9589	0.3535
Baseline	0.0441	0.0837	0.9597	0.9591	0.3445
Baseline+MEM	0.0423	0.0831	0.9601	0.9594	0.2724
Baseline+USK	0.0417	0.0825	0.9607	0.9598	0.2463
Base+Refine	**0.0401**	**0.0817**	**0.9610**	**0.9602**	**0.2001**

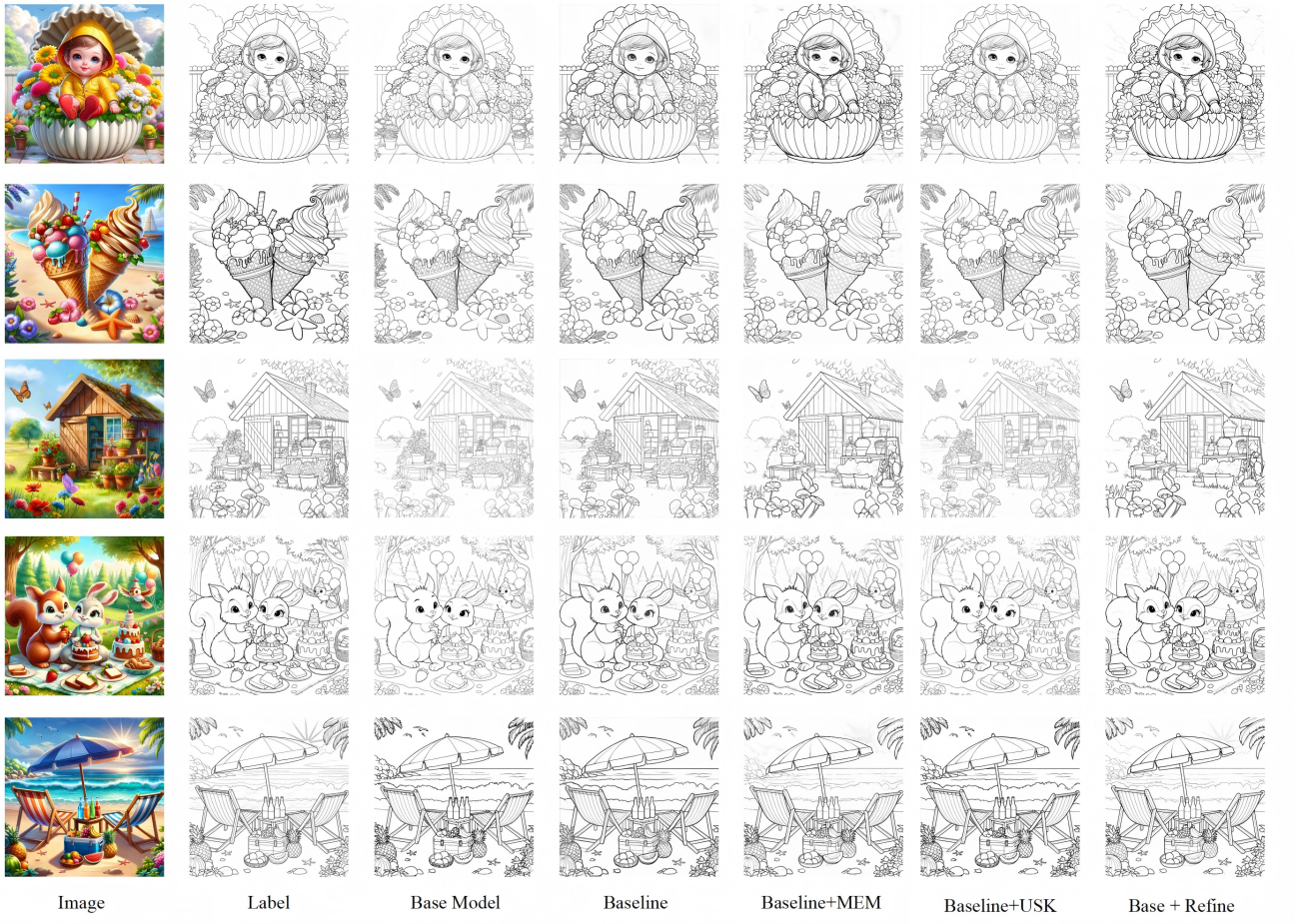

Figure 5. Visualization of ablation experiment

0.9602, and 0.2001, respectively. These results demonstrate the overall effectiveness of the complete framework in improving pixel-wise accuracy, structural consistency, and perceptual quality.

4.6. Generalization experiment

To assess the generalization capability of DiCo-Net, we further conducted experiments on the Anime Sketch Colorization Pair dataset, which is characterized by relatively low annotation quality. The line-art in this dataset differs markedly from the LINE-2K dataset used in our primary study, particularly in line-thickness variability and heterogeneous hand-drawn styles. Training and validation strictly followed the dataset's official partitioning, and the training strategy as well as all hyperparameters were kept consistent with the main experiments. Fig. 6 reports the cross-dataset comparison. Relative to competing methods, DiCo-Net adapts effectively to the stylistic characteristics of the Anime Sketch Colorization Pair dataset, maintaining high structural integrity and line continuity. As shown in Fig. 7, benefiting from DiCo-Net's feature-processing mechanism and noise-suppression strategy, the generated line-art exhibits visually smoother strokes and lower overall noise than the ground-truth annotations, further demonstrating stable generalization across datasets.

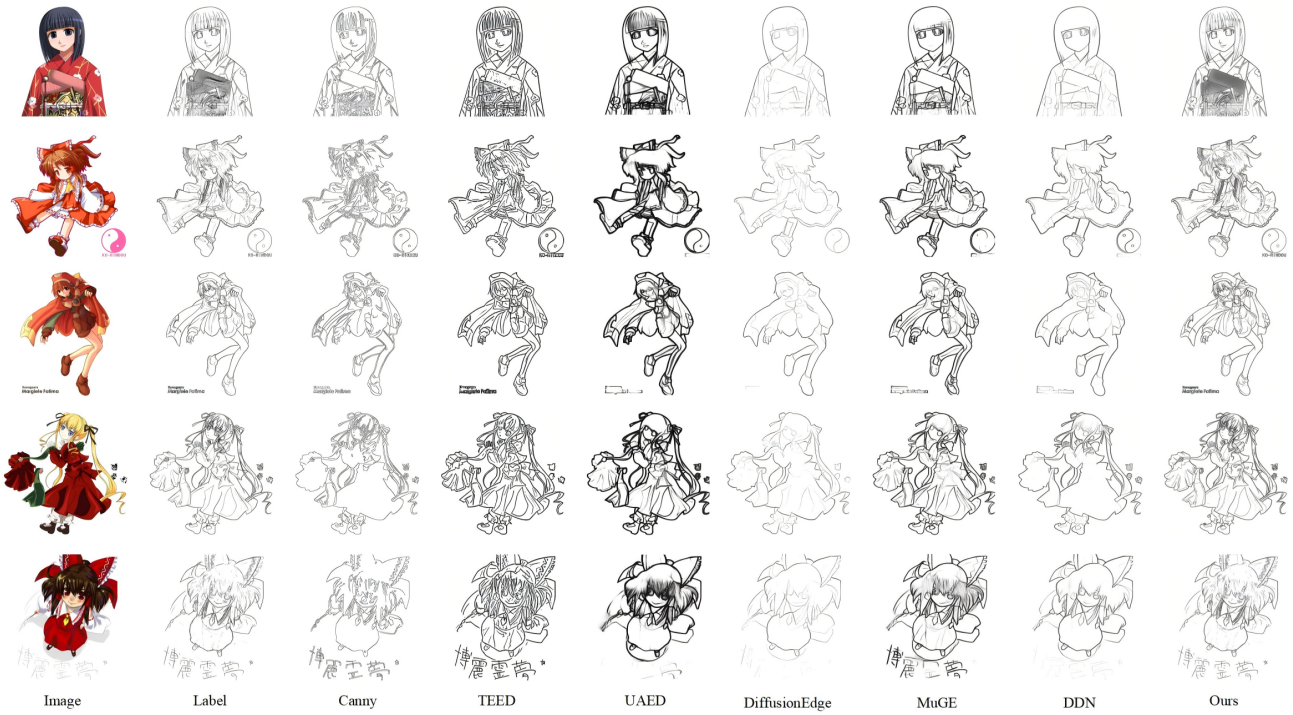

Figure 6. Comparison of generalization results across different models

Table 3. User study results.

Model	Aesthetic Mean Score↑	Artistic Structure Mean Score↑	Aesthetic Std↓	Artistic Structure Std↓
Canny	1.63	1.47	0.32	0.29
TEED	2.19	2.38	0.27	0.19
UAED	2.46	2.25	0.33	0.27
DiffusionEdge	2.95	3.12	0.22	0.20
MuGE	3.37	3.24	0.28	0.21
DDN	3.84	3.58	0.27	0.19
Ours	**4.41**	**4.23**	**0.22**	**0.18**

4.7. User Research

To assess the effectiveness of line-art generation from the perspective of users' subjective and authentic experiences, a user study was conducted. The evaluation framework comprises two principal dimensions: first, Aesthetic Quality, which quantifies the overall visual impact and aesthetic appeal of the generated line-art; and second, Artistic Structure, which evaluates artistic expressiveness and structural integrity in terms of semantic coherence, line smoothness, and detail fidelity. In this experiment, 50 images were randomly sampled from the test set. Each image presented the output of the proposed method alongside those from 2–3 comparative approaches. To eliminate potential bias, all results were shown to participants under blind testing conditions. A total of 30 participants were recruited, including both art professionals with drawing experience and general users, ensuring diversity and representativeness in subjective assessment. For each image, participants rated both Aesthetic Quality and Artistic Structure on a 5-point Likert scale, where a score of 1 indicated "very unappealing/structurally chaotic," and a score of 5 indicated "very appealing/structurally coherent."

As summarized in Table 3, our method significantly outperforms all baselines in subjective ratings. For aesthetic appeal, the proposed approach attains an average score of 4.41, exceeding DDN's 3.84 and MuGE's 3.37, indicating the highest user preference in overall visual impact and appeal. For artistic structure, our method achieves an average of 4.23, surpassing all competitors (DDN: 3.58, MuGE: 3.24), which reflects superior semantic coherence, line fluidity, and detail depiction aligned with expectations for professional line-art. Moreover, the standard deviations for aesthetic appeal and artistic structure are 0.22 and 0.18, respectively, indicating concentrated and consistent user feed-

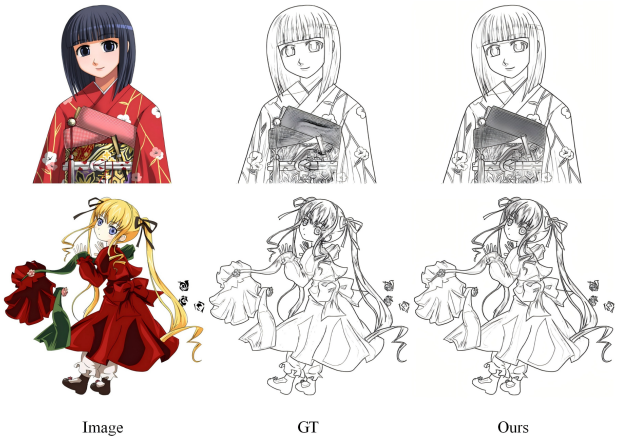

Figure 7. Generalization experiment results

back with high subjective agreement. Overall, the proposed method attains both high mean scores and low variance across dimensions, producing line-art that better satisfies professional requirements for semantic completeness and detail presentation.

5. Conclusion

This paper proposes DiCo-Net, a two-stage line-art extraction framework that combines diffusion-based generation with convolutional refinement. Through the multimodal enhancement module (MEM) and the adaptive convolutional attention mechanism (USK), the proposed method enables the joint fusion of texture and depth cues while modeling fine-grained stroke structures and multi-scale structural information, leading to improved detail fidelity and semantic coherence. By incorporating multi-scale supervision via a binary cross-entropy loss, the model strikes a balance between fine-grained prediction accuracy and structural integrity across multiple scales. Experimental results on the LINE-2K dataset show that DiCo-Net consistently outperforms existing approaches in terms of edge continuity, fine-line responsiveness, background artifact suppression, and overall perceptual quality, producing results that are visually closer to professionally drawn line art.

Despite these improvements, the proposed method still relies on computationally intensive diffusion models during inference, and certain computational redundancies exist among network components. Future work will focus on developing lightweight architectures through model distillation and optimizing existing modules to reduce computational overhead and improve engineering efficiency. In addition, we plan to investigate more robust line-art extraction from low-quality or noisy real-world images, as well as to explore unsupervised and self-supervised learning strategies to reduce reliance on high-quality paired data and enhance the model's generalization capability in complex scenarios.

Acknowledgement

This research was funded by the "AI Coloring Content Generation and Line Art Drawing Project" (Project No. NKRI-YW-202507001). We would like to thank Xi'an Button Software Technology Co., Ltd. for its valuable support throughout the research process, particularly in the construction of a high-precision hand-drawn line art dataset. The company's professional art team provided high-quality rendering and detailed annotations, which ensured the reliability of the experimental data and supported the smooth progression of this research.

References

[1] J. Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 2009. 1, 2

[2] B. Cetinkaya, S. Kalkan, and E. Akbas. Ranked: Addressing imbalance and uncertainty in edge detection using ranking-based losses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3239–3249, 2024. 2

[3] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10850–10869, 2023. 2, 3

[4] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang. Bi-directional cascade network for perceptual edge detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3828–3837, 2019. 2

[5] L. Huan, N. Xue, X. Zheng, W. He, J. Gong, and G.-S. Xia. Unmixing convolutional features for crisp edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6602–6609, 2021. 1, 2

[6] J. Jie, Y. Guo, G. Wu, J. Wu, and B. Hua. Edgenat: transformer for efficient edge detection. *arXiv preprint arXiv:2408.10527*, 2024. 3

[7] B. F. Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne, T. Dockhorn, J. English, Z. English, P. Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 2, 3

[8] M. Li, D. Chen, and S. Liu. Beta network for boundary detection under nondeterministic labels. *Knowledge-Based Systems*, 266:110389, 2023. 2

[9] X. Li, W. Wang, X. Hu, and J. Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 510–519, 2019. 3

[10] Y. Li, X. S. Poma, Y. Xi, G. Li, C. Yang, Q. Xiao, Y. Bai, and Z. Li. A doubly decoupled network for edge detection. *Neurocomputing*, 624:129442, 2025. 2

[11] Y. Li, T. Yao, Y. Pan, and T. Mei. Contextual transformer networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):1489–1500, 2022. 3

[12] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai. Richer convolutional features for edge detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3000–3009, 2017. 1, 2

[13] X. Liufu, C. Tan, X. Lin, Y. Qi, J. Li, and J.-F. Hu. Sauge: Taming sam for uncertainty-aligned multi-granularity edge detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5766–5774, 2025. 3

[14] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217, 1980. 1, 2

[15] X. S. Poma, E. Riba, and A. Sappa. Dense extreme inception network: Towards a robust cnn model for edge detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1923–1932, 2020. 2

[16] M. Pu, Y. Huang, Y. Liu, Q. Guan, and H. Ling. Edter: Edge detection with transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1402–1412, 2022. 1, 2

[17] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020. 3

[18] I. Sobel, G. Feldman, et al. A 3x3 isotropic gradient operator for image processing. *a talk at the Stanford Artificial Project in*, 1968:271–272, 1968. 1, 2

[19] X. Soria, Y. Li, M. Rouhani, and A. D. Sappa. Tiny and efficient model for the edge detection generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1364–1373, 2023. 2

[20] X. Soria, G. Pomboza-Junez, and A. D. Sappa. Ldc: Lightweight dense cnn for edge detection. *IEEE Access*, 10:68281–68290, 2022. 2

[21] Z. Su, W. Liu, Z. Yu, D. Hu, Q. Liao, Q. Tian, M. Pietikäinen, and L. Liu. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5117–5127, 2021. 1, 2

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[23] S. Xie and Z. Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 1, 2

[24] W. Xuan, S. Huang, J. Liu, and B. Du. Fcl-net: Towards accurate edge detection via fine-scale corrective learning. *Neural Networks*, 145:248–259, 2022. 2

[25] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 2

[26] Y. Ye, K. Xu, Y. Huang, R. Yi, and Z. Cai. Diffusionedge: Diffusion probabilistic model for crisp edge detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 6675–6683, 2024. 2, 3

[27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[28] C. Zhou, Y. Huang, M. Pu, Q. Guan, R. Deng, and H. Ling. Muge: Multiple granularity edge detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25952–25962, 2024. 1, 3

[29] C. Zhou, Y. Huang, M. Pu, Q. Guan, L. Huang, and H. Ling. The treasure beneath multiple annotations: An uncertainty-aware edge detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15507–15517, 2023. 2

[30] C. Zhou, Y. Huang, M. Xiang, J. Ren, H. Ling, and J. Zhang. Generative edge detection with stable diffusion. *arXiv preprint arXiv:2410.03080*, 2024. 2, 3