SceneDiff: Generative Scene-Level Image Retrieval with Text and Sketch Using Diffusion Models

Ran Zuo[1,2⋆] , **Haoxiang Hu**[1,2⋆] , **Xiaoming Deng**[1,2†] , **Cangjun Gao**[1,2] , **Zhengming Zhang**[1,2] , **Yu-Kun Lai**[3] , **Cuixia Ma**[1,2,4†] , **Yong-Jin Liu**[5†] , **Hongan Wang**[1,2]

[1]Beijing Key Laboratory of Human-Computer Interaction, Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Cardiff University
[4]Key Laboratory of System Software and State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences
[5]Tsinghua University

Abstract

Jointly using text and sketch for scene-level image retrieval utilizes the complementary between text and sketch to describe the fine-grained scene content and retrieve the target image, which plays a pivotal role in accurate image retrieval. Existing methods directly fuse the features of sketch and text and thus suffer from the bottleneck of limited utilization for crucial semantic and structural information, leading to inaccurate matching with images. In this paper, we propose SceneDiff, a novel retrieval network that leverages a pre-trained diffusion model to establish a shared generative latent space, enabling a joint latent representation learning for both sketch and text features and precise alignment with the corresponding image. Specifically, we encode text, sketch and image features, and project them into the diffusion-based share space, conditioning the denoising process on sketch and text features to generate latent fusion features, while employing the pre-trained autoencoder for latent image features. Within this space, we introduce the content-aware feature transformation module to reconcile encoded sketch and image features with the diffusion latent space's dimensional requirements and preserve their visual content information. Then we augment the representation capability of the generated latent fusion features by integrating multiple samplings with partition attention, and utilize contrastive learning to align both direct fusion features and generated latent fusion features with corresponding image representations. Our method outperforms the state-of-the-art works through extensive experiments, providing a novel insight into the related retrieval field.

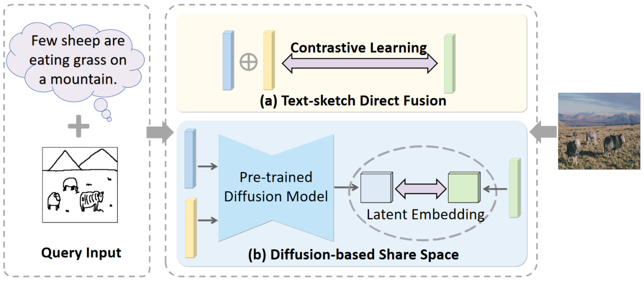

Few sheep are eating grass on a mountain.
Query Input
Contrastive Learning
(a) Text-sketch Direct Fusion
Pre-trained Diffusion Model
Latent Embedding
(b) Diffusion-based Share Space

Figure 1: Illustration of generative scene-level image retrieval with text and sketch as queries. (a) Conventional methods directly fuse the features of sketch and text and align it with the image for retrieval. Besides the direct feature fusion, our method also (b) generates the latent fusion feature via the diffusion model and aligns it with the latent image feature in the generative share space, which enables effective feature fusion between sketch and text and feature matching for retrieval.

1 Introduction

Scene-level text and sketch-based image retrieval (Scene-level TSBIR) aims to depict the image scene content with the collaboration of text and hand-drawn sketches, which exploits the strengths of both semantic and appearance description in text (e.g., object categories, texture) and objects' structural attributes depicted in sketches (e.g., object layout, relative size, and shape). By integrating features from text and sketch, it establishes a robust match with target images, significantly improving the accuracy of image retrieval.

Existing works [Sangkloy *et al.*, 2022; Chowdhury *et al.*, 2023] often rely on direct feature fusion strategies for sketch and text features, such as element-wise summation [Sangkloy *et al.*, 2022] or cross-attention [Chowdhury *et al.*, 2023], and align the fused feature with image through contrastive learning. However, these approaches can only maintain insufficient crucial information coverage which overlooks specific details of sketches and text, and they can enable limited feature interaction which lacks full utilization of individual sketch and text features to learn a comprehensive joint representation, and hinders accurate correspondence between two modalities and image features. Therefore, we introduce a generative model into the retrieval framework, which constructs the scene-level correlation between sketches and text and strengthens the semantic and visual alignment between generated fusion features and image features in the shared generative feature space.

Recent years have witnessed a paradigm shift in generative modeling with the emergence of diffusion models. These models have transcended their initial application in image generation and are now demonstrating significant utility across diverse domains, including image semantic segmentation [Amit *et al.*, 2021; Baranchuk *et al.*, 2021; Brempong *et al.*, 2022], object detection [Chen *et al.*, 2023] and localization [Zhao *et al.*, 2023c], representation learning for text and images [Zhao *et al.*, 2023b], cross-modal video retrieval based on text [Jin *et al.*, 2023], and video localization [Li *et al.*, 2023; Zhao *et al.*, 2023a]. Among these advancements, Stable Diffusion [Rombach *et al.*, 2022] stands out for its ability to generate images conditioned on text and support for incorporating additional structural information, such as sketches, skeletons, or semantic segmentation maps, alongside text as joint conditional inputs to the model [Zhang *et al.*, 2023a; Mou *et al.*, 2023]. This enables the progressive denoising process to reveal not only the relationship between text and sketches but also their intricate mapping onto the generated images. Inspired by the generative models, we design new method of sketch and text feature fusion and feature alignment between fused text/sketch features and images.

In this paper, we propose SceneDiff, a novel method to leverage the diffusion-based generative model in the scene-level TSBIR task. We first perform feature encoding on the text, sketches and images. Then the diffusion-based generator is introduced to utilize the sketch and text features as conditions for the denoising process, which allows the sketch and text features to be mapped into the image latent space, resulting in generated latent fusion features. Simultaneously, image features are also mapped to the generative space to obtain latent image features. In order to deal with the issue that the encoded sketch and image features are not consistent with the original input of the diffusion model, we design a content-aware feature transformation module, which projects the sketch and image features into the generative space while preserving their visual details. In order to enhance the representation robustness of the generated latent fusion features, we propose a generated feature enhancement module, which adopts multiple sampling strategies and integrates these generated features via the partition attention mechanism. Finally, we utilize two-level contrastive learning to perform feature alignment, between the directly fused sketch and text features with image features after feature encoding (shown in Figure 1(a)), and between the generated latent fusion features and the latent image features in the diffusion-based share space (shown in Figure 1(b)). SceneDiff incorporates a specific diffusion-based retrieval architecture to learn the joint representation between sketch and text, facilitating robust interaction between the two modalities. This model further enhances the feature representation capabilities of encoders by employing the generative model to train dedicated encoders for each modality. Extensive experiments demonstrate the proposed method significantly outperforms existing approaches in terms of retrieval accuracy.

The main contributions of this work are listed as follows:

- We are the first to propose the generative retrieval framework for scene-level image retrieval with text and sketch, which leverages the diffusion model to optimize feature fusion of text and sketch as well as its alignment with the image in the diffusion-based share space.
- We introduce a content-aware feature transformation module that provides a pathway for mapping sketch and image features onto the generative space of the pretrained diffusion model, while preserving their inherent visual information.
- We augment the generated latent fusion features through a novel partition-based attention mechanism with multiple samplings in latent space, which can generate more representative and robust latent fusion representation for feature matching with image.
- Experiments on multiple datasets for the scene-level TSBIR task demonstrate that our proposed generative retrieval method achieves state-of-the-art performance.

2 Related Work

2.1 Image Retrieval with Text and Sketch

The utilization of text and sketch as queries for image retrieval has been extensively studied in the literature. Previous works [Dey *et al.*, 2018; Han and Schlangen, 2017; Song *et al.*, 2017] have explored the benefits of training models using both sketch and text inputs to improve the image retrieval performance for each modality separately. These works typically combine the retrieval results obtained from text and sketch during test time, without considering the feature-level correlation of both modalities.

Recent methods have been developed to integrate text and image features to achieve comprehensive feature fusion and alignment with the image [Sangkloy *et al.*, 2022; Radford *et al.*, 2021]. Specifically, Task-former [Sangkloy *et al.*, 2022] extends CLIP [Radford *et al.*, 2021] to support an additional sketch input. It adopts a late fusion strategy to combine the encoded sketch and text features and uses contrastive learning for the fused sketch-text features and image features. A multi-label classification loss and a caption generation loss are added to enhance the model's ability to recognize the semantics of each modality. SceneTrilogy [Chowdhury *et al.*, 2023] introduces a novel method to disentangle three modalities into modality-specific and modality-agnostic features. It extracts modality-agnostic features of sketches, text and images for image retrieval and fuses sketch and text features via a cross-attention mechanism. These two methods directly fuse the features of sketches and text and enhance the alignment with image features through contrastive learning. However, they may result in the loss of crucial information inherent in each modality, leading to suboptimal utilization of their features and impairing the feature alignment with images. We adopt the retrieval framework by extending CLIP as our baseline motivated by Task-former [Sangkloy *et al.*, 2022] and integrate a generative model into the framework to address the challenges mentioned above.

2.2 Diffusion Models in Cross-domain Retrieval

Generative models play a significant role in cross-modal retrieval tasks, as they can enhance the diversity of samples and establish sample correspondences in the generative space. Among generative models, diffusion models stand out for their success in retrieval tasks like text-to-image [Zhao *et al.*, 2023b] and text-to-video [Jin *et al.*, 2023; Li *et al.*, 2023; Zhao *et al.*, 2023a], demonstrably enhancing retrieval accuracy. RLEG [Zhao *et al.*, 2023b] leverages a pre-trained diffusion model to sample semantically similar data, enhancing text and image feature representations for more robust feature alignment and effective generalization to unseen data. Its multi-sampling strategy leads to significant improvements in feature alignment by comparing multiple generated representations. The well-trained feature encoder can then be effectively leveraged for downstream text-to-image retrieval tasks, demonstrating strong performance in various retrieval benchmarks. DiffusionRet [Jin *et al.*, 2023] utilizes the diffusion model to model the joint distribution of text and video data. By incorporating the progressive denoising process, it effectively uncovers the semantic relationships between text and video modalities. It exhibits remarkable generalization capabilities, effectively handling out-of-domain samples. Motivated by these works, we leverage diffusion models to enhance the feature fusion of sketch and text and align them with image features, realizing the improvement of retrieval accuracy and the optimization of encoders for comprehensive modality representation learning.

3 Method

As illustrated in Figure 2, the proposed retrieval network SceneDiff consists of two major components. It first extends the network architecture of CLIP [Radford *et al.*, 2021] as the basic framework, which incorporates dedicated encoders to extract pertinent features from sketch, text, and image, and then learns the correspondence between directly fused sketch-text features with image features through contrastive learning. Subsequently, it leverages a pre-trained diffusion model to project sketch and text features as conditions, precisely guiding the denoising process of the noisy image's latent feature. It obtains the generated latent fusion features by mapping the sketch and text features into the latent space of the image domain and then incorporates an additional contrastive loss to learn the correspondence between the generated latent fusion features and native latent image features. Within the diffusion-based share space, we introduce a new content-aware feature transformation module (CAFT), which transforms sketch and image features to align with the input dimension of the pre-trained diffusion model and preserves their inherent visual information simultaneously. To further augment the representation capacity of the generated latent fusion features, we incorporate a generated feature enhancement module (GFEM), which employs a partition attention mechanism to integrate the most representative features from multiple generated samples.

3.1 Preliminary: Model Construction

Given the robust feature representation capabilities of CLIP through pre-training on large-scale datasets, we leverage an extended CLIP architecture as the foundational retrieval framework similar to [Sangkloy *et al.*, 2022]. We adopt two CLIP image encoders for sketch features F_S and image features F_I, respectively, and a CLIP text encoder for text features F_T. Then we introduce the diffusion model-based retrieval framework to map features of sketch, text, and image to a generative space and achieve efficient feature fusion and alignment. Specifically, we employ Stable Diffusion (SD) [Rombach *et al.*, 2022] as the generative framework due to its capacity for text-driven image synthesis and refer T2I-Adapter [Mou *et al.*, 2023] to simultaneously incorporate sketch for the denoising process guidance.

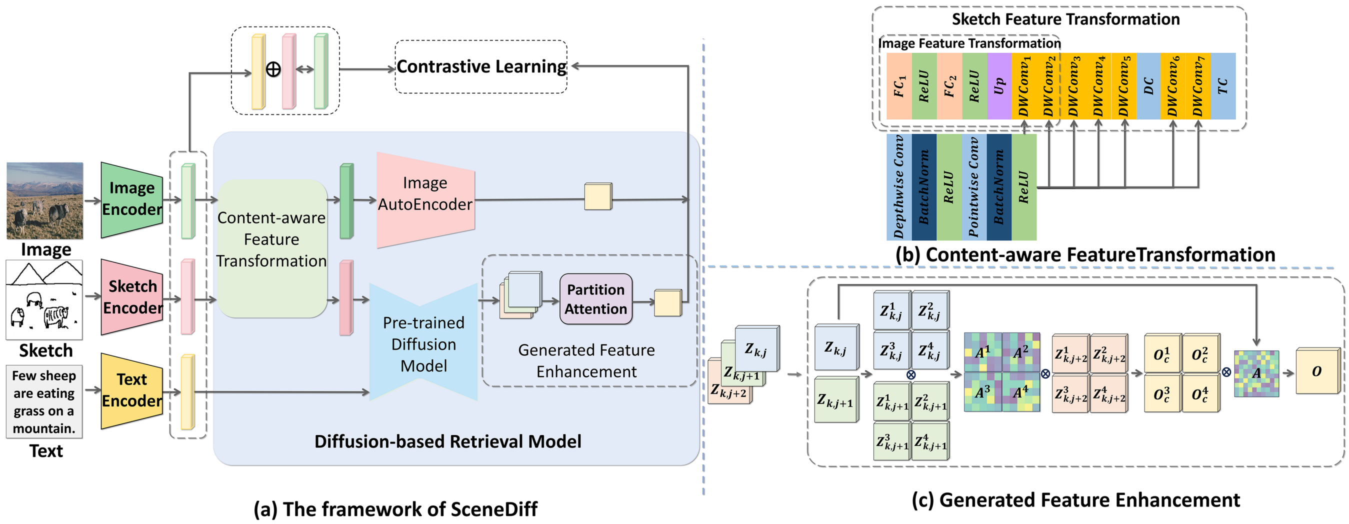
SD model [Rombach *et al.*, 2022] operates within a two-stage framework for image generation. In the first stage, it adopts a pre-trained autoencoder to encode image x into the latent feature z_0 and gradually adds noise to obtain z_t that follows a random Gaussian distribution. In the denoising phase, the latent feature z_t is fed into a UNet network to denoise into z_0 with t steps. The denoising process is guided by the text condition c, which is encoded through a frozen CLIP text encoder. The noise-free latent embedding z_0 is decoded by the autoencoder to generate the final image. T2I-Adapter [Mou *et al.*, 2023] supports adding additional structural condition c_s into the pre-trained SD model, where sketch passes through an Adapter module to extract features and downsample to produce multi-scale condition features $F_{c_s} = \{F_{c_s}^i\}(i = 1, ..., 4)$. They have the same dimensions with intermediate features $F_e = \{F_e^i\}(i = 1, ..., 4)$ of the denoising UNet encoder and incorporate into UNet via a layer-wise additive mechanism:

$$F_e^i = F_e^i + F_{c_s}^i, \quad i = 1, ..., 4 \tag{1}$$

Within the SceneDiff framework, we leverage the text feature F_T as the text condition c and construct multi-scale sketch features $\{F_S^i\}(i = 1, ..., 4)$ as the structural condition c_s, feeding both into the pre-trained SD model similar to that of Mou et al. [Mou *et al.*, 2023]. Through the denoising process, the sketch and text features are mapped into the image domain and obtain the generated latent fusion feature $z_I{}'$. The image feature F_I is simultaneously projected into the generative space through the pre-trained autoencoder and obtains the latent image feature z_I. Subsequently, a two-level contrastive loss is employed to model the correspondence between these representations, fostering an abundant shared space that bridges the visual and textual modalities.

3.2 Content-aware Feature Transformation

Given that the SceneDiff model leverages a CLIP text encoder to extract text feature F_T, compatible with the text condition encoding methodology employed in the SD model, we directly utilize F_T as text condition c to guide the denoising process. However, it is incongruous between sketch and image features because the pre-trained SD model and T2I-Adapter operate on pixel-level features for both image and


Contrastive Learning
Image Encoder
Image
Sketch Encoder
Sketch
Few sheep are eating grass on a mountain.
Text
Text Encoder
Content-aware Feature Transformation
Image AutoEncoder
Pre-trained Diffusion Model
Partition Attention
Generated Feature Enhancement
Diffusion-based Retrieval Model
Sketch Feature Transformation
Image Feature Transformation
FC_1
ReLU
FC_2
ReLU
Up
$DWConv_1$
$DWConv_2$
$DWConv_3$
$DWConv_4$
$DWConv_5$
DC
$DWConv_6$
$DWConv_7$
TC
Depthwise Conv
BatchNorm
ReLU
Pointwise Conv
BatchNorm
ReLU
$z_{k,j}$
$z_{k,j+1}$
$z_{k,j+2}$
$z^1_{k,j}$ $z^2_{k,j}$ $z^3_{k,j}$ $z^4_{k,j}$
$z^1_{k,j+1}$ $z^2_{k,j+1}$ $z^3_{k,j+1}$ $z^4_{k,j+1}$
A^1 A^2 A^3 A^4
$z^1_{k,j+2}$ $z^2_{k,j+2}$ $z^3_{k,j+2}$ $z^4_{k,j+2}$
o^1_c o^2_c o^3_c o^4_c
A
o

(a) The framework of SceneDiff

(b) Content-aware FeatureTransformation

(c) Generated Feature Enhancement

Figure 2: (a) Framework of SceneDiff to realize scene-level TSBIR with the diffusion model. We project encoded sketch-text features and image features into the diffusion-based generative latent space, and employ two-level contrastive learning for matching, i.e. matching between directly fused sketch-text features (element-wise sum) with image features, and matching between generated latent fusion features with latent image features. In order to enhance the diffusion model's fitness for the retrieval task, we introduce two modules (i.e. (b) and (c)). (b) Content-aware Feature Transformation module transforms the sketch and image features to match the input dimension requirement of the diffusion model without sacrificing visual fidelity; (c) Generated Feature Enhancement module utilizes a partition attention mechanism to integrate multiple generated samples to enhance the representation capacity of generated features.

sketch, while SceneDiff provides the token-level representations F_S and F_I, which necessitates a new feature transformation module to bridge the gap and achieve compatibility. Therefore, we propose a content-aware feature transformation module (CAFT), which transforms significant features of images and sketches into diverse sizes, ensuring their precise alignment with the pre-trained SD model's dimensional constraints and preserving their specific content.

Image Feature Transformation To effectively distill global and local features when modulating the dimension of image feature F_I, we employ a sequential architecture comprising a two-layer fully connected (FC) module followed by the depthwise separable convolution (DWConv). The architecture is shown in "Image Feature Transformation" part of Figure 2 (b). The initial 2 FC layers integrate the global information of F_I. Subsequently, DWConv extracts local features with a computationally efficient combination of depthwise and pointwise convolutions while reducing model complexity. They re-extract the comprehensive representation from the encoded image feature and reshape the feature dimension to facilitate the subsequent process.

Sketch Feature Transformation To ensure dimensional congruity between sketch features and intermediate features $F_e = \{F_e^i\}(i = 1, ..., 4)$ within the denoising UNet encoder, we transform the sketch feature F_S into multi-scale $F_S = \{F_S^i\}(i = 1, ..., 4)$ through the transformation module in "Sketch Feature Transformation" part of Figure 2 (b). Beyond the combination of 2 FC layers and 7 DWConv layers for multi-scale sketch feature extraction, we leverage the dilated convolution (DC) behind the 5th DWConv layer to integrate multi-scale features from the sketch to further capture global context without additional computational costs. To mitigate the spatial information attrition during deep feature extraction, we integrate transposed convolution (TC) behind the 7th DWConv layer to help recover the sketch's fine-grained visual details. Finally, we obtain $F_S = \{F_S^i\}(i = 1, ..., 4)$ from the 2nd and 4th DWConv layers, the DC layer and TC layer (see more details in Supplementary), and add to the UNet encoder intermediate features F_e indicated in Eq. 1.

3.3 Generated Feature Enhancement

The inherent stochasticity of the pre-trained diffusion model presents challenges for negative exemplar and unrepresentative feature generation when relying on a single sample. To address this challenge, we introduce the generated feature enhancement module (GFEM) which conducts multiple incompletely denoised samplings and employs a partition attention mechanism to integrate them. This module exhibits greater reliability and comprehensiveness of the generated features, thereby enhancing the model's robustness and generalizability to improve performance in downstream retrieval tasks.

Multiple Samplings with Incomplete Denoising We generate n samples from the pre-trained SD model to augment the generated latent features. Existing research [Zhao *et al.*, 2023b; Jin *et al.*, 2023] suggests that incorporating the DDIM sampling strategy [Song *et al.*, 2020] in retrieval tasks yields efficient feature representations with reduced sampling steps, which facilitates accelerated sampling and improves the overall efficiency of the retrieval process. Therefore, we take the result in the k-th denoising step as representation for each sampling, resulting in a set of generated latent features $\{z_{k,j}\}(j = 1, ..., n)$.

Partition Attention with Multiple Samplings To address the potential for noise contamination within global features arising from the direct fusion of incomplete denoising latent features, we propose a partition attention mechanism based on cross-attention [Vaswani *et al.*, 2017], which partitions the feature space into localized regions, enabling a more refined and noise-resilient fusion process.

As illustrated in Figure 2 (c), the generated n latent features are divided into $n/3$ groups, each of which contains 3 samples $\{z_{k,j}, z_{k,j+1}, z_{k,j+2}\} \in \mathbb{R}^{C\times H\times W}$. Then, the features within each group undergo a channel-wise partition process, yielding an ensemble of b distinct blocks, and a cross-attention mechanism is applied to each block b_i as follows:

$$A^i = \frac{z^i_{k,j}{z^i_{k,j+1}}^T}{\sqrt{d}} \in \mathbb{R}^{C/b\times H\times W}, (i = 1, .., b) \tag{2}$$

$$O^i = softmax(A^i) z^i_{k,j+2} \tag{3}$$

where d is the dimension of $z^i_{k,j}$ and $z^i_{k,j+1}$, and C/b, H, and W are the channel, height, and width of the attention matrix A^i, respectively. O^i is the feature produced by the i-th partitioned block. Then we concatenate the partitioned block features $\{O^i\}$ to get the feature O_c:

$$O_c = concat([O^1, ..O^b]) \tag{4}$$

Then we apply a global attention map to update the feature O_c ensuring the global feature distribution consistency:

$$O = A \times O_c = \frac{z_{k,j}{z_{k,j+1}}^T}{\sqrt{d}} \times O_c, A \in \mathbb{R}^{C\times H\times W} \tag{5}$$

The partition attention mechanism is applied to each group of n features, and we get the generated latent fusion feature $z_I{}'$ by the sum of the feature set $O = \{O_1, ...O_{n/3}\}$.

3.4 Loss Function

The learning objective of SceneDiff consists of two main components: contrastive learning between the direct fusion features and image features, and contrastive learning between the generated latent fusion features and latent image features. This two-level contrastive learning approach further strengthens the model's ability to establish adequate correspondences across modalities, ultimately enhancing the overall effectiveness of the retrieval process.

Contrastive Learning with Direct Fusion Feature Given a set of text-sketch-image pairs $\{F^i_T, F^i_S, F^i_I\}^N_{i=1}$, we first sum the sketch features F_S with text features F_T to obtain the fused features F_f, and then apply the InfoNCE Loss [Oord *et al.*, 2018] to a batch of fusion-image feature pairs $\{F^i_f, F^i_I\}^B_{i=1}$, aligning the direct fusion feature F^i_f with the image feature F^i_I through fuse-to-image and image-to-fuse contrastive losses:

$$\mathcal{L}_{f2i} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(F^i_f \cdot F^{i^+}_I/\tau\right)}{\sum_{j=1}^{N}\exp\left(F^i_f \cdot F^{j^-}_I/\tau\right)} \tag{6}$$

$$\mathcal{L}_{i2f} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(F^i_I \cdot F^{i^+}_f/\tau\right)}{\sum_{j=1}^{N}\exp\left(F^i_I \cdot F^{j^-}_f/\tau\right)} \tag{7}$$

where $F^{i^+}_I$, $F^{j^-}_I$ is the respective positive and negative image sample to the fusion feature F^i_f, and $F^{i^+}_f$, $F^{j^-}_f$ is the positive and negative fusion sample to the image feature F^i_I, τ is the temperature parameter to control the scale of pairwise cosine similarities.

Channel Attention-based Contrastive Learning with Generated Latent Fusion Feature Beyond employing contrastive learning between the directly fused sketch-text features and image features, we additionally leverage contrastive learning between the generated latent fusion feature $z_I{}'$ and the latent image feature z_I. To strengthen the alignment between $z_I{}'$ and z_I, we incorporate the channel-attention mechanism to allocate attention to the most pertinent channels, ensuring accuracy and consistency within the latent representations. The generated latent fusion feature $z_I{}'$ is processed through the channel attention mechanism as follows:

$$z_I{}' = z_I{}' \odot \sigma\left(W_2 \cdot ReLU\left(W_1 \cdot AvgPool(z_I{}')\right)\right) \tag{8}$$

where $AvgPool$ stands for global average pooling, W_1 and W_2 denote the weight matrices of the fully connected layers, $ReLU$ and σ are the activation functions, and $\odot$ signifies element-wise multiplication. The latent image feature z_I is updated by applying the same operation in Eq. 8 by replacing z'_I with z_I.

After updating the generated latent fusion feature $z_I{}'$ and latent image feature z_I with the channel-attention mechanism, we proceed to apply the InfoNCE loss as follows:

$$\mathcal{L}_{z_I 2 z_I{}'} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(z^i_I \cdot (z_I{}')^{i^+}/\tau\right)}{\sum_{j=1}^{N}\exp\left(z^i_I \cdot (z_I{}')^{j^-}/\tau\right)} \tag{9}$$

$$\mathcal{L}_{z_I{}' 2 z_I} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left((z_I{}')^i \cdot z^{i^+}_I/\tau\right)}{\sum_{j=1}^{N}\exp\left((z_I{}')^i \cdot z^{j^-}_I/\tau\right)} \tag{10}$$

The total loss function is:

$$\mathcal{L}_{total} = \lambda_1\left(\mathcal{L}_{f2i} + \mathcal{L}_{i2f}\right) + \lambda_2\left(\mathcal{L}_{z_I 2 z_I{}'} + \mathcal{L}_{z_I{}' 2 z_I}\right) \tag{11}$$

where λ_1 and λ_2 are loss weight parameters to determine the loss distribution.

4 Experiments

4.1 Implementation Details

We initialize the encoders for sketches, text, and images in the SceneDiff model with the publicly available CLIP model (ViT-B/16). Then we construct the diffusion-based retrieval framework by utilizing the pre-trained SD model with version 1.4, along with its associated pre-trained autoencoder. The parameters are set as follows: the number of samplings n is 3, the number of sampling steps k is 2, λ_1 and λ_2 is 1 and 0.1 respectively. All experiments are conducted on one NVIDIA A100 80G GPU with learning rate 1e-6 and batch size 4. More details are listed in the supplementary.

Query Input	Method	SFSD			FS-COCO			SketchyCOCO		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Text	CLIP	16.1	33.57	43.03	9.63	22.52	30.53	25.24	54.29	65.71
	Fine-tuned CLIP	19.32	41.55	56.49	13.42	28.05	36.72	29.66	60.15	71.82
	SceneDiff (w Text)	**21.08**	**44.97**	**58.82**	**14.02**	**28.56**	**37.16**	**31.42**	**59.26**	**77.34**
Sketch	SketchyScene	60.01	72.83	80.12	22.85	40.9	51.19	27.51	54.23	74.1
	SceneSketcher	69.58	82.29	86.4	/	/	/	31.9	66.71	86.2
	SceneDiff (w Sketch)	**71.8**	**82.41**	**88.76**	**25.17**	**45.93**	**55.93**	**34.29**	**69.05**	**81.43**
Text&Sketch	TASK-former	78.52	93.13	95.63	40.27	62.65	75.86	38.04	58.18	69.24
	SceneTrilogy	/	/	/	25.7	/	55.2	39.5	/	88.7
	SceneDiff	**85.1**	**96.43**	**98.85**	**46.37**	**67.71**	**78.52**	**61.02**	**83.33**	**92.76**

Table 1: Comparision of the retrieval performance between SceneDiff and SOTA image retrieval methods on SketchyCOCO, FS-COCO, SFSD datasets.

4.2 Datasets and Evaluation Metrics

We utilize three scene-level sketch-image datasets with textual descriptions for our retrieval task: (1) **SketchyCOCO** [Gao *et al.*, 2020] contains 14,081 synthetic sketch-image pairs, where images are selected from the MS-COCO dataset [Lin *et al.*, 2014] with associated text. Given that most sketches include less than one foreground instance, we adopt the filtering approach of Scene Sketcher [Liu *et al.*, 2020] to select 1,015 pairs for training and 210 for testing. (2) **FS-COCO** [Chowdhury *et al.*, 2022] is a hand-drawn sketch-image dataset with textual descriptions of sketches, which includes 7,000/3,000 train/test pairs. (3) **SFSD** [Zhang *et al.*, 2023b] is another hand-drawn sketch-image dataset containing 12,115 pairs. The text in this dataset corresponds to the associated image, which is also selected from MS-COCO [Lin *et al.*, 2014]. We divide the dataset into 8,480/3,635 train/test pairs. Given scene-level TSBIR, we use R@K where K is set to 1, 5, and 10 for evaluation, representing the percentage of queries for which the target images are present within the top K retrieval results.

4.3 Comparison with State-of-the-art

We compare SceneDiff with the following state-of-the-art retrieval methods. For text-based image retrieval (TBIR) methods, **CLIP** [Radford *et al.*, 2021] is pre-trained on large-scale datasets and well-suited for text-driven image retrieval. **Fine-tuned CLIP** fine-tunes the parameters of CLIP on different sketch-image datasets. For scene-level sketch-based image retrieval (scene-level SBIR) methods, **SketchyScene** [Zou *et al.*, 2018] combines InceptionV3 [Szegedy *et al.*, 2016] and the triplet ranking network [Yu *et al.*, 2016] together to retrieve the target image. **SceneSketcher** [Liu *et al.*, 2020] introduces a novel graph-based approach for fine-grained scene-level SBIR, which constructs the graph with object instances serving as nodes and updates graph features via the Graph Convolutional Networks (GCN) [Kipf and Welling, 2016]. For text and sketch-based image retrieval (TSBIR) methods, **TASK-former** [Sangkloy *et al.*, 2022] extends the CLIP architecture to incorporate sketches, text, and images, and employs multiple losses to align the sum of sketch and text features with image features. **SceneTrilogy** [Chowdhury *et al.*, 2023] adopts feature disentanglement to extract modality-agnostic features of text, sketch, and image. It combines sketch and text features via cross-attention for TSBIR.

Table 1 presents the comparison results on datasets of SketchyCOCO, FS-COCO, and SFSD. Due to the original codes for SceneTrilogy [Chowdhury *et al.*, 2023] have not been released, we directly cite its retrieval results on SketchyCOCO and FS-COCO. Furthermore, the lack of instance-level annotation in FS-COCO prevents SceneSketcher from performing retrieval on this dataset. It is obvious that SceneDiff consistently outperforms existing TSBIR methods across diverse datasets, demonstrating its broad applicability and potential for real-world TSBIR deployment. The model's specific retrieval results are shown in Figure 3. Even when using a single sketch or text as queries, it surpasses both TBIR and scene-level SBIR methods, showcasing the remarkable versatility of diffusion models in enhancing performance across various cross-domain retrieval tasks.

4.4 Ablation Study

To evaluate the impact of each component on the SceneDiff model, we conduct an ablation study on the SFSD dataset: (1) **Baseline**: our baseline encodes the sketch and image features with the CLIP image encoder and text features with the CLIP text encoder, then fuses text and sketch features through summation and aligns with image features through contrastive learning. (2) **Baseline + Diff.**: we incorporate a pre-trained SD model into the retrieval framework, leveraging its latent space for cross-modal fusion and alignment. Then we transform sketch and image features to the input specification of the SD model via FC layers and upsampling. Subsequently, sketch features undergo the pre-trained Adapter module [Mou *et al.*, 2023] to obtain multi-scale structural condition features F_{c_s} and collaborate with text features to guide the denoising process for generated latent fusion features. In parallel, image features traverse the pre-trained autoencoder to attain latent representations. Two-level contrastive learning is employed for direct fusion features and image features, as well as generated latent fusion features and latent image features. (3) **Baseline+Diff.+CAFT**: the content-aware feature transformation module (CAFT) is proposed as a replacement for the original transformation method and the pre-trained Adapter module. (4) **Baseline+Diff.+GFEM**: the generated feature

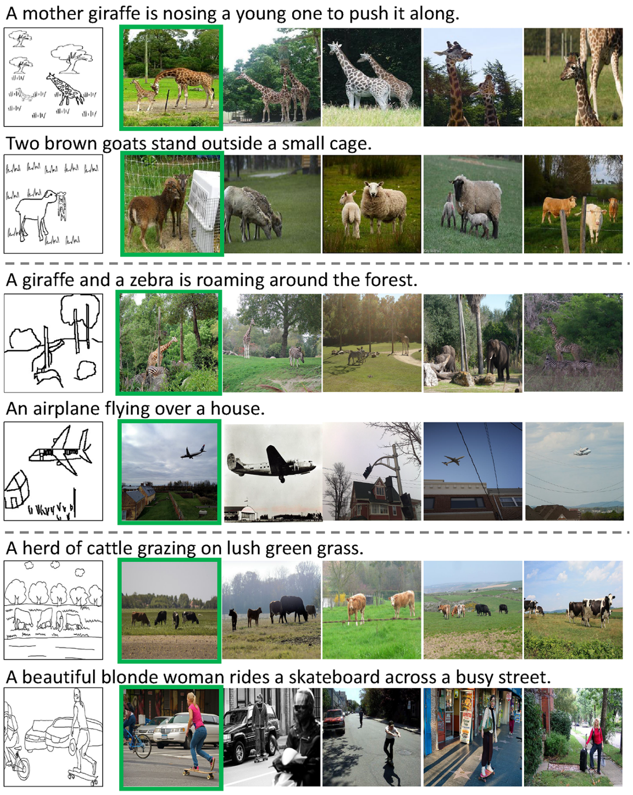

A mother giraffe is nosing a young one to push it along.
Two brown goats stand outside a small cage.
A giraffe and a zebra is roaming around the forest.
An airplane flying over a house.
A herd of cattle grazing on lush green grass.
A beautiful blonde woman rides a skateboard across a busy street.

Figure 3: Top-5 retrieval results of SceneDiff on the SketchyCOCO, FS-COCO, and SFSD datasets. The true matches are highlighted with green rectangles.

enhancement module (GFEM) is added to improve the robustness and representativeness of the generated latent fusion features. (5) **Baseline+Diff.+Channel-Att.**: the channel-attention (Channel-Att.) is used to augment the representation consistency of generated latent fusion features and latent image features before contrastive learning. (6) **Full Model**: we assemble the complete retrieval model by integrating all modules, collectively improving the model's retrieval performance. As presented in Table 2, each component contributes to the retrieval accuracy improvement. Direct integration of the pre-trained SD model proves suboptimal for the retrieval framework. CAFT, GFEM, and Channel-Att. are introduced to enhance its performance, with CAFT demonstrating the most significant impact. The best retrieval performance is achieved through the combination of these components.

4.5 Retrieval with Incomplete Query Input

Considering the frequent incompleteness of input text or sketches in practical applications, we conduct the ablation study to evaluate the SceneDiff model's robustness for incomplete query inputs. We randomly mask 0%-50% words of text and 0%-50% content of the sketch. The model trained on complete query inputs is evaluated under two conditions: (1) complete sketches with incomplete text, and (2) incomplete sketches with complete text. As illustrated in Table 3, the model retains satisfactory retrieval accuracy with missing text, demonstrating its applicability in situations where textual inputs might be incomplete. Compared to text, sketches exert a greater influence on the overall retrieval performance. Therefore, prioritizing sketch integrity in practical applications is crucial for optimal model effectiveness.

Baseline	Diff.	CAFT	GFEM	Channel-Att.	R@1
✓					78.06
✓	✓				79.63
✓	✓	✓			83.56
✓	✓		✓		80.41
✓	✓			✓	81.97
✓	✓	✓	✓	✓	85.1

Table 2: The ablation study on components of the SceneDiff model using the SFSD dataset. "Diff.": the diffusion-based retrieval network, "CAFT": the content-aware feature transformation module, "GFEM": the generated feature enhancement module, "Channel-Att.": the channel attention mechanism.

Type	Incompleteness	R@1	R@5	R@10
Text	0%	85.1	96.43	98.85
	10%	83.7	96.2	98.57
	20%	81.67	95.07	97.37
	30%	79.43	93.67	96.93
	40%	76.77	92.43	96.17
	50%	74.96	90.83	94.47
Sketch	0%	85.1	96.43	98.85
	10%	67.97	89.03	94.23
	20%	50.3	75.27	83.73
	30%	45.03	59.73	70.97
	40%	33.03	45.7	57.13
	50%	24.7	36.0	46.4

Table 3: The retrieval results on the SFSD dataset with incomplete text or sketch inputs, including (1) complete sketch and incomplete text with 0%-50% content missing as query input, (2) complete text and incomplete sketch with 0%-50% content missing as query input.

5 Conclusion

In this paper, we propose a novel scene-level TSBIR retrieval framework that incorporates a pre-trained SD model to enhance the fusion of sketch and text and alignment with image. The model commences with separate encoding of sketch, text, and image features. Subsequently, a content-aware feature transformation module projects sketch and image features into the diffusion-based share space. Conditioned on sketch-text features, the latent fusion features are generated through a denoising process and bolstered for robust representation via the generated feature enhancement module. Lastly, contrastive learning establishes correspondences between directly fused sketch-text features and image features, as well as generated latent fusion features and latent image features. The proposed method achieves state-of-the-art performance for scene-level TSBIR, offers a fresh direction for research on this important task, and holds the potential for generalization to related retrieval tasks in future studies.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62272447 and 62332019, and Science and Technology Innovation Key R&D Program of Chongqing (CSTB2023TIAD-STX0027).

Contribution Statement

Ran Zuo and Haoxiang Hu are equal contributions. Xiaoming Deng, Cuixia Ma and Yong-Jin Liu are corresponding authors for this study.

References

[Amit *et al.*, 2021] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

[Baranchuk *et al.*, 2021] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[Brempong *et al.*, 2022] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4175–4186, 2022.

[Chen *et al.*, 2023] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843, 2023.

[Chowdhury *et al.*, 2022] Pinaki Nath Chowdhury, Aneeshan Sain, Ayan Kumar Bhunia, Tao Xiang, Yulia Gryaditskaya, and Yi-Zhe Song. Fs-coco: Towards understanding of freehand sketches of common objects in context. In *European Conference on Computer Vision*, pages 253–270. Springer, 2022.

[Chowdhury *et al.*, 2023] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Scenetrilogy: On human scene-sketch and its complementarity with photo and text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10972–10983, 2023.

[Dey *et al.*, 2018] Sounak Dey, Anjan Dutta, Suman K Ghosh, Ernest Valveny, Josep Lladós, and Umapada Pal. Learning cross-modal deep embeddings for multi-object image retrieval using text and sketch. In *2018 24th international conference on pattern recognition (ICPR)*, pages 916–921. IEEE, 2018.

[Gao *et al.*, 2020] Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. Sketchycoco: Image generation from freehand scene sketches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5174–5183, 2020.

[Han and Schlangen, 2017] Ting Han and David Schlangen. Draw and tell: Multimodal descriptions outperform verbal-or sketch-only descriptions in an image retrieval task. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 361–365, 2017.

[Jin *et al.*, 2023] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. Diffusionret: Generative text-video retrieval with diffusion model. *arXiv preprint arXiv:2303.09867*, 2023.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[Li *et al.*, 2023] Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. Momentdiff: Generative video moment retrieval from random to real. *arXiv preprint arXiv:2307.02869*, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[Liu *et al.*, 2020] Fang Liu, Changqing Zou, Xiaoming Deng, Ran Zuo, Yu-Kun Lai, Cuixia Ma, Yong-Jin Liu, and Hongan Wang. Scenesketcher: Fine-grained image retrieval with scene sketches. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 718–734. Springer, 2020.

[Mou *et al.*, 2023] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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

[Sangkloy *et al.*, 2022] Patsorn Sangkloy, Wittawat Jitkrittum, Diyi Yang, and James Hays. A sketch is worth a thousand words: Image retrieval with text and sketch. In *Eu-*

ropean Conference on Computer Vision, pages 251–267. Springer, 2022.

[Song *et al.*, 2017] Jifei Song, Yi-Zhe Song, Tony Xiang, and Timothy M Hospedales. Fine-grained image retrieval: the text/sketch input dilemma. In *BMVC*, volume 2, page 7, 2017.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Yu *et al.*, 2016] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 799–807, 2016.

[Zhang *et al.*, 2023a] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[Zhang *et al.*, 2023b] Zhengming Zhang, Xiaoming Deng, Jinyao Li, Yukun Lai, Cuixia Ma, Yongjin Liu, and Hongan Wang. Stroke-based semantic segmentation for scene-level free-hand sketches. *The Visual Computer*, 39(12):6309–6321, 2023.

[Zhao *et al.*, 2023a] Henghao Zhao, Kevin Qinghong Lin, Rui Yan, and Zechao Li. Diffusionvmr: Diffusion model for video moment retrieval. *arXiv preprint arXiv:2308.15109*, 2023.

[Zhao *et al.*, 2023b] Liming Zhao, Kecheng Zheng, Yun Zheng, Deli Zhao, and Jingren Zhou. Rleg: vision-language representation learning with diffusion-based embedding generation. In *International Conference on Machine Learning*, pages 42247–42258. PMLR, 2023.

[Zhao *et al.*, 2023c] Yuzhong Zhao, Qixiang Ye, Weijia Wu, Chunhua Shen, and Fang Wan. Generative prompt model for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6351–6361, 2023.

[Zou *et al.*, 2018] Changqing Zou, Qian Yu, Ruofei Du, Haoran Mo, Yi-Zhe Song, Tao Xiang, Chengying Gao, Baoquan Chen, and Hao Zhang. Sketchyscene: Richly-annotated scene sketches. In *Proceedings of the european conference on computer vision (ECCV)*, pages 421–436, 2018.