algorithms

MDPI

Article

GDUI: Guided Diffusion Model for Unlabeled Images

Xuanyuan Xie and Jieyu Zhao *

Mobile Network Application Technology Laboratory, School of Information Science and Engineering, Ningbo University, 818 Fenghua Road, Ningbo 315211, China; 2111082089@nbu.edu.cn
* Correspondence: zhao_jieyu@nbu.edu.cn

Abstract: The diffusion model has made progress in the field of image synthesis, especially in the area of conditional image synthesis. However, this improvement is highly dependent on large annotated datasets. To tackle this challenge, we present the Guided Diffusion model for Unlabeled Images (GDUI) framework in this article. It utilizes the inherent feature similarity and semantic differences in the data, as well as the downstream transferability of Contrastive Language-Image Pretraining (CLIP), to guide the diffusion model in generating high-quality images. We design two semantic-aware algorithms, namely, the pseudo-label-matching algorithm and label-matching refinement algorithm, to match the clustering results with the true semantic information and provide more accurate guidance for the diffusion model. First, GDUI encodes the image into a semantically meaningful latent vector through clustering. Then, pseudo-label matching is used to complete the matching of the true semantic information of the image. Finally, the label-matching refinement algorithm is used to adjust the irrelevant semantic information in the data, thereby improving the quality of the guided diffusion model image generation. Our experiments on labeled datasets show that GDUI outperforms diffusion models without any guidance and significantly reduces the gap between it and models guided by ground-truth labels.

Keywords: image synthesis; guided diffusion; semantic aware; pseudo-label matching

Citation: Xie, X.; Zhao, J. GDUI: Guided Diffusion Model for Unlabeled Images. *Algorithms* **2024**, *17*, 125. https://doi.org/10.3390/a17030125

Academic Editor: Arslan Munir

Received: 24 February 2024
Revised: 8 March 2024
Accepted: 14 March 2024
Published: 18 March 2024

1. Introduction

Visual computing is rapidly advancing with the emergence of generative artificial intelligence (AI). This revolutionary technology enables unprecedented capabilities in generating, editing, and reconstructing images, videos, and 3D scenes [1]. Diffusion models, as powerful generative models, have made significant contributions in image synthesis, 3D reconstruction, and semantic segmentation. The accompanying issue is that in these fields, especially in image synthesis, a large labeled dataset [2,3] is required to ensure the diversity and fidelity of the generated images. The fidelity of the diffusion model can be enhanced by utilizing class labels to guide the reverse denoising process [4,5]. To further improve the image generation process, classifier guidance [5] or classifier-free guidance [6] can be utilized to provide better guidance. However, these conditions and guiding methods rely on datasets that require ground-truth annotations, which can be challenging to obtain in many fields, such as remote sensing [7] and medical imaging [8]. Therefore, the objective of this article is to better align unlabeled images with real class information and apply them to the guided diffusion model. To further improve the image generation process, classifier guidance [5] or classifier-free guidance [6] can be utilized to provide better guidance. However, these conditions and guiding methods rely on datasets that require ground-truth annotations, which can be challenging to obtain in many fields, such as remote sensing [7] and medical imaging [8]. Therefore, the objective of this article is to better align unlabeled images with real class information and apply them to the guided diffusion model.

Specifically, the requirement for annotated datasets for successful image synthesis should be reduced or even eliminated. Currently, the use of self-supervised guided diffu-

sion models [9,10] has shown significant improvements in guiding effectiveness. These methods do not require any labeled information and solve pretext tasks at the image level. However, since the guidance information produced by this self-supervision is often disconnected from the information of the real categories, we are unable to generate images of specific real categories during image generation.

Therefore, we propose the GDUI framework, which uses guided diffusion to generate images of specified real categories without requiring specific labeled datasets. The GDUI framework can be divided into three stages. In the first stage of GDUI, the image is encoded into a semantically meaningful latent vector using a self-supervised clustering framework, and the diffusion pseudo-label-matching algorithm is used to further map the semantic information. In the second stage, GDUI uses the label-matching refinement algorithm to filter out irrelevant semantic information. In the third stage, the conditional diffusion model uses the matched real semantic information of the image as guidance to generate the image.

We summarize the contributions as follows. (1) We propose GDUI, which achieves the task of guided image generation based on real categories on unlabeled datasets, through the similarity of the data itself and the downstream transferability of CLIP [11]. (2) We design a pseudo-label-matching algorithm that completes the matching of the real semantic information of the image based on image similarity to help the model better accomplish the guided generation of specified real categories. (3) We design a label-matching refinement algorithm to adjust irrelevant semantic information of the data. (4) The experimental results demonstrate that, on unlabeled datasets, the fidelity and diversity of images generated by guided generation with specified categories are superior to unconditionally generated images, and even surpass those generated by using ground-truth labels.

2. Related Work

2.1. Conditional Diffusion Models

In recent years, conditional diffusion models have shown remarkable achievements in domains such as image synthesis [5,12], text-to-image [3,13], image restoration [14,15], and more. The conditional diffusion models have been extending their conditional information beyond class labels [5] to include other modalities, such as textual descriptions [16,17] and semantic segmentation [18]. It provides a high degree of controllability through various guiding mechanisms, including classifier-guided [5,19], CLIP-guided [20], and classifier-free diffusion [6,12]. DALL-E2 [3] and GLIDE [21] can generate high-quality images based on large-scale text–image datasets. The aforementioned models have all achieved significant results, but they all relied on paired text–image datasets for training. Some models, such as KNN diffusion [22] and retrieval-augmented diffusion [23], utilized nearest neighbor samples during training to reduce their dependence on labeled datasets. LAFITE [24] has once again demonstrated the feasibility of using pre-trained CLIP [11] for text-to-image generation. Compared to these works, we achieve high-quality image generation by better aligning image and label information.

2.2. Deep Clustering

Deep clustering methods utilize the representation ability of deep neural networks and have shown superiority over traditional clustering algorithms. Early deep clustering methods mostly combined stacked auto-encoders (SAE) [25] with traditional clustering algorithms, such as spectral clustering [26], Gaussian mixture model [27,28], and subspace clustering [29,30]. However, because the pixel-wise reconstruction loss of SAE tends to over-emphasize low-level features [31], it leads to a loss of object-level semantic information, resulting in poor clustering performance for images with complex content. Recently, there have been clustering methods [32,33] that combine self-supervised and contrastive learning. However, these approaches often overlook the semantic differences between clusters and focus only on the similarity between instances, which can limit the clustering performance [31]. The semantic pseudo-labeling-based image clustering method,

SPICE [31], was proposed to effectively balance the similarity among instances and the semantic discrepancies between clusters, thereby improving clustering performance. Our research leverages the significant potential of self-supervised methods in downstream tasks [34]. Specifically, we focus on combining self-supervised learning with semantic pseudo-labeling in an image clustering approach for image generation. By incorporating this approach, we provide valuable guidance signals to diffusion models, empowering them to generate high-quality images.

2.3. Vision and Language (VL) Models

In recent years, vision–language models have made significant progress in various fields, such as visual question answering [35], image–text retrieval [36,37], and more. While early tasks mainly focused on specific domains, there is an increasing demand for vision–language models with greater generality in practical applications. Therefore, recent works such as CLIP [11] and ALIGN [38] were built on a general framework for visual and language representation. CLIP [11] trained language and image information jointly by minimizing the distance between corresponding image and text embeddings, achieving zero-shot performance on downstream tasks. Due to the impressive performance of vision–language models [39], our research focuses on improving the alignment between images and semantic information using these models. The objective is to provide diffusion models with more accurate guidance signals.

3. Methodology

Given an unlabeled image dataset $\mathfrak{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ consisting of N images, where $\boldsymbol{x}_i$ represents the i-th unlabeled image, our objective is to associate $\mathfrak{X}$ with the true labels of K categories. In other words, our goal is to match each image $\boldsymbol{x}_i$ with its corresponding true label y_i^g, where y_i^g represents the true label of the i-th image. Subsequently, we employ diffusion models and classifiers to process the labeled images and generate high-quality images guided by their respective categories. To accomplish the aforementioned objectives, we propose the GDUI. The overall flow of the proposed GDUI for unlabeled images manipulation is illustrated in Figure 1. First, the input unlabeled images $\mathfrak{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ are clustered into K classes. Then, the pseudo-label-matching algorithm is used to transform the image set with pseudo-labels into a set of images with true labels. Second, we fine-tune the labels of the images using the label-matching refinement algorithm. Third, we optimize guided diffusion using labeled images matched by the label-matching refinement algorithm. In the following subsections, we will first provide a brief background on diffusion models, followed by a detailed dissection of the individual modules.

3.1. Preliminary

Diffusion probabilistic models [40,41] are a class of latent variable models that involve both a forward diffusion process and a reverse diffusion process. The forward process of diffusion model is a Markov chain where data are gradually corrupted with Gaussian noise based on a variance schedule $\beta_1, \ldots, \beta_T$:

$$q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0) := \prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) \tag{1}$$

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) := \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right) \tag{2}$$

where $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_T$ are latent variables of the same dimension, and $\boldsymbol{x}_0$ follows the distribution $q(\boldsymbol{x}_0)$. The inverse process of the diffusion model, denoted as $p_\theta(\boldsymbol{x}_{0:T})$, is defined as a Markov chain with learned Gaussian transitions:

$$p_\theta(\boldsymbol{x}_{0:T}) := p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) \tag{3}$$

$$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) := \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)\boldsymbol{I}) \tag{4}$$

where $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ can be represented as a linear combination of $\boldsymbol{x}_T$ and a noise predictor $\epsilon_\theta(\boldsymbol{x}_t, t)$, the variance $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ is fixed to a known constant typically.

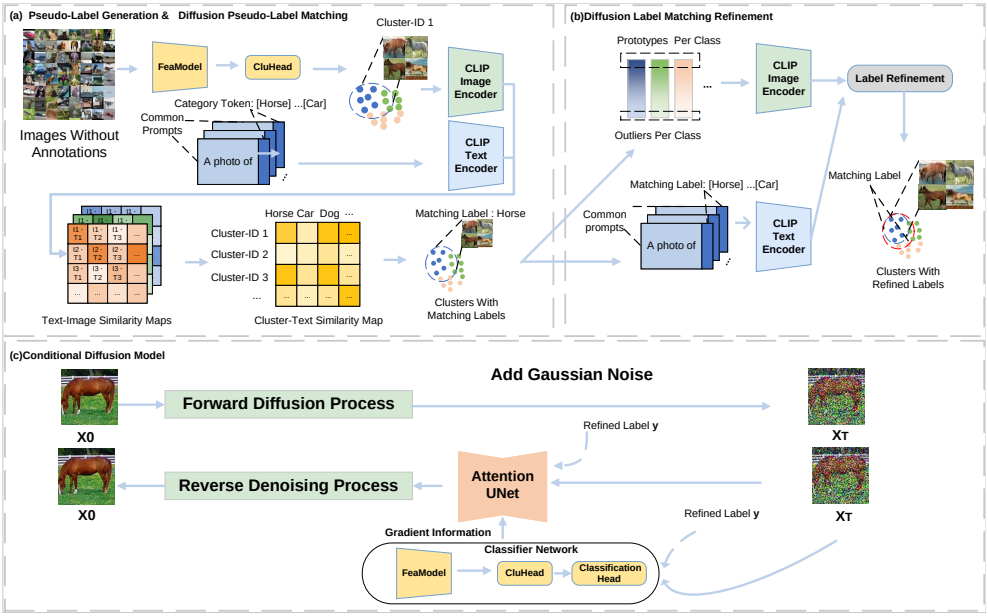


Figure 1. Illustration of the GDUI framework. (**a**) Encode images into semantically meaningful latent vectors and map semantic information. (**b**) Filter out irrelevant information. (**c**) Generate images based on matched real semantic information.

The quality of samples can be optimized by the following parameterized and simplified objective:

$$L_{simple}(\theta) := \mathbb{E}_{t,\boldsymbol{x}_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2\right] \tag{5}$$

Here, t is uniformly distributed between 1 and T. It is defined $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

Compared to unconditional diffusion models, conditional diffusion models can generate images specified by conditions. The classifier-guided [5,42] sampling method demonstrates that the gradient $\nabla_{\boldsymbol{x}_t} \log p_\phi(y \mid \boldsymbol{x}_t)$ of a classifier can guide conditional diffusion models to generate higher-quality samples with a specified class y.

3.2. Pseudo-Label Generation

To extend the classifier guidance technique to unlabeled images, here we adopt a deep clustering approach for the unsupervised learning of visual features [43] to cluster the samples and generate synthetic labels. We adopt the SPICE [31] framework that divides network training into three stages. First, there are two branches in which two different random transformations of the same image are taken as inputs. Each branch includes a feature model and a projection head. Given two transformations x' and x'' of an image x, the outputs of the two branches are represented as z^+ and z, respectively. The parameters of the feature model $\mathcal{F}$ and projection head $\mathcal{P}$ are optimized by the following loss function:

$$\mathcal{L}_{fea} = -\log\left(\frac{\exp\left(z^{\mathrm{T}} z^+/\tau\right)}{\sum_{i=1}^{N_q}\exp\left(z^{\mathrm{T}} z_i^-/\tau\right) + \exp\left(z^{\mathrm{T}} z^+/\tau\right)}\right) \tag{6}$$

where z_i^- is the negative sample and τ is the temperature. The finally optimized feature model parameters are denoted as $\boldsymbol{\theta}_{\mathcal{F}}^s$.

In the second stage, given the feature model parameters $\boldsymbol{\theta}_{\mathcal{F}}^{s}$ and the unlabeled images $\mathfrak{X}$, the goal is to separately optimize the parameters $\boldsymbol{\theta}_{\mathcal{C}}$ of the clustering head $\mathcal{C}$ in order to predict cluster labels $\{y_i^s\}$. The optimization of parameters $\boldsymbol{\theta}_{\mathcal{C}}$ is performed within the EM framework, where the cluster labels $\{y_i^s\}$ are obtained given $\boldsymbol{\theta}_{\mathcal{C}}$ in the expectation (E) step, and then in the maximization (M) step, the parameters $\boldsymbol{\theta}_{\mathcal{C}}$ are optimized upon obtaining the cluster labels $\{y_i^s\}$.

In the third stage, the feature model $\mathcal{F}$ and clustering head $\mathcal{C}$ are jointly optimized. After obtaining the embedding features $\{f_i\}_{i=1}^{N}$ and cluster labels $\{y_i^s\}$ corresponding to the images $\mathfrak{X}$ in the first two stages, a subset of reliable samples $\mathfrak{X}^r$ is selected as:

$$\mathfrak{X}^r = \{(\boldsymbol{x}_i, y_i^s) \mid r_i > \sigma_r, \forall i = 1, 2, \ldots, N\} \tag{7}$$

where r_i is the semantically consistent ratio of the sample $\boldsymbol{x}_i$ and σ_r denotes the confidence threshold for $\mathfrak{X}^r$. The semantically consistent ratio r_i of the sample $\boldsymbol{x}_i$ is defined as:

$$r_i = \frac{1}{N_s} \sum_{y \in \mathfrak{L}_i} \mathbb{1}(y = y_i^s) \tag{8}$$

where N_s represents the number of samples that are closest to the sample $\boldsymbol{x}_i$ based on the cosine similarity between their embedding features, and $\mathfrak{L}_i$ represents the corresponding labels of these N_s samples. The jointly trained network optimizes the parameters $\boldsymbol{\theta}_{\mathcal{F}}$ and $\boldsymbol{\theta}_{\mathcal{C}}$ using the following loss function:

$$\mathcal{L}_{joint} = \frac{1}{L} \sum_{i=1}^{L} \mathcal{L}_{ce}(y_i^s, \mathcal{C}(\mathcal{F}(\alpha(\boldsymbol{x}_i); \boldsymbol{\theta}_{\mathcal{F}}); \boldsymbol{\theta}_{\mathcal{C}})) + \frac{1}{U} \sum_{j=1}^{U} \mathbb{1}\left(y_j^u \geq 0\right) \mathcal{L}_{ce}\left(y_j^u, \mathcal{C}\left(\mathcal{F}(\beta(\boldsymbol{x}_j); \boldsymbol{\theta}_{\mathcal{F}}); \boldsymbol{\theta}_{\mathcal{C}}\right)\right) \tag{9}$$

where the first part is calculated using L reliable samples $\left(x_i, y_i^s\right)$ from $\mathfrak{X}^r$, and the second part is calculated using U pseudo-labeled samples $\left(x_i, y_i^u\right)$ with pseudo-labels y_j^u. These pseudo-labels y_j^u are assigned to the classes predicted by the network with the highest probability and exceeding a certain threshold. α and β respectively denote the operators for weak and strong transformations on the input image. $\mathcal{L}_{ce}$ is the cross-entropy loss function.

After three stages of clustering, the input unlabeled images $\mathfrak{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ are divided into K clusters with clustering labels $\{y_i^s\}$. The probability matrix $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_N]^{\mathrm{T}} \in \mathbb{R}^{N \times K}$ is generated for the image set for each clusters. The probability matrix $\boldsymbol{P}$ represents the probabilities of each image belonging to specific clusters.

3.3. Diffusion Pseudo-Label Matching

Based on the obtained cluster labels $\{y_i^s\}$ of $\mathfrak{X}$ over K clusters, our goal is to match them with the ground-truth labels to guide target generation in the GDUI model. In unsupervised situations, we do not have ground truth to match against. To address the challenge of matching the ground-truth labels with the obtained cluster labels and to ensure a globally attentive alignment, we adopt the principles of the Stable Marriage Algorithm (SMA) [44] for the overall matching strategy. This approach emphasizes the importance of considering global information in the matching process. To address this issue, we propose the pseudo-label-matching algorithm, which leverages the zero-shot capability of CLIP to achieve bilateral matching between the clustering labels and the ground truth. Given the clustering probability matrix $\boldsymbol{P}$ and K clusters with cluster labels $\{y_i^s\}$, the top confident samples are selected as the clustering prototypes for each cluster.

To illustrate the process, we take the m-th cluster as an example. We define the m-th cluster as:

$$\mathfrak{X}_m = \{(\boldsymbol{x}_i, y_i^s) \mid y_i^s = m, \forall i = 1, 2, \ldots, N\} \tag{10}$$

The top samples are selected as:

$$\mathfrak{X}_m^{top} = \{(\boldsymbol{x}_i, y_i^s) \mid i \in \text{argtopk}(\boldsymbol{P}_{j,m}, N_{top}), \boldsymbol{x}_j \in \mathfrak{X}_m\} \tag{11}$$

where the argtopk function argtopk$\left(\boldsymbol{P}_{j,m}, N_{top}\right)$ returns the indices of the top N_{top} highest-scoring samples in the m-th cluster.

Using CLIP, zero-shot classification is performed on the samples in $\mathfrak{X}_m^{top}$ with respect to the provided ground-truth label set $\mathcal{Y}^g = \left\{y_i^g\right\}_{i=1}^K$. The classes with the highest classification probability are then selected as the labels for the samples. We obtain $\boldsymbol{p}_i^c$ as the CLIP classification probability for the m-th cluster by calculating the proportion of each class in the N_{top} samples. This ensures that the priority of each class in the clustering result is directly proportional to its probability such that higher probabilities correspond to higher priorities. Then, based on the previously obtained K clusters, the CLIP priority matrix for the K clusters, $\boldsymbol{P}^c = \left[\boldsymbol{p}_1^c, \boldsymbol{p}_2^c, \ldots, \boldsymbol{p}_K^c\right]^{\mathrm{T}} \in \mathbb{R}^{K\times K}$, is constructed.

A cluster, for example, the m-th cluster $\mathfrak{X}_m$, is selected from the unmatched clusters $\mathfrak{U} = \{\mathfrak{X}_i\}_{i=1}^K$ that have not been matched with the ground-truth label set $\mathcal{Y}^g$, and the highest priority class y_i^g, represented as:

$$y_i^g = \text{argmax}\, \boldsymbol{P}_{m,j}^c, j \in \mathcal{Y}_m^u \tag{12}$$

where $\mathcal{Y}_m^u$ represents the ground truth of the m-th cluster that has not been requested for label matching, and $\boldsymbol{P}_{m,j}^c$ denotes the element in the m-th row and column j of matrix $\boldsymbol{P}^c$, where j is the index of the element in the $\mathcal{Y}_m^u$. Then, argmax $\boldsymbol{P}_{m,j}^c$ returns the index of the highest priority class among all unmatched ground truths for the m-th cluster. If the class y_i^g has not been assigned, then it is assigned to the current m-th cluster. Otherwise, its priority is compared with the already assigned clusters. The cluster with a lower priority is added back to the unmatched clustering set $\mathfrak{U}$, while the one with a higher priority is matched with class y_i^g. Until all clusters are matched, we can obtain each cluster and its corresponding label, denoted by $\mathfrak{M} = \left\{\left(\mathfrak{X}_i, y_i^g\right)\right\}_{i=1}^K$, where each label corresponds to the true class.

The above process for diffusion pseudo-label matching is summarized in Algorithm 1.

Algorithm 1 Pseudo-label matching

Input: Unmatched clusters $\mathfrak{U} = \{\mathfrak{X}_i\}_{i=1}^K$

Output: Matched clusters $\mathfrak{M} = \left\{\left(\mathfrak{X}_i, y_i^g\right)\right\}_{i=1}^K$

for $i = 1, 2, \ldots, K$ **do**
  Select i-th cluster $\mathfrak{X}_i$ from $\mathfrak{U}$ with Equation (10);
  Select N_{top} top confident samples $\mathfrak{X}_i^{top}$ from $\mathfrak{X}_i$ with Equation (11);
  Compute class proportions $\boldsymbol{p}_i^c$ based on highest classification probability of each $\boldsymbol{x}_i \in \mathfrak{X}_i^{top}$ using CLIP;
end for
Generate priority matrix $\boldsymbol{P}^c = \left[\boldsymbol{p}_1^c, \boldsymbol{p}_2^c, \ldots, \boldsymbol{p}_K^c\right]^{\mathrm{T}}$;
while $\mathfrak{U} \neq \varnothing$ **do**
  Pick a cluster $\mathfrak{X}_m$ from $\mathfrak{U}$;
  Select the highest priority true label y_i^g among unrequested matches with Equation (12);
  if y_i^g has not been assigned **then**
    Assign y_i^g to cluster $\mathfrak{X}_m$;
  else
    y_i^g has been assigned to cluster $\mathfrak{X}_k$;
    Assign y_i^g to the cluster with higher priority between $\mathfrak{X}_m$and $\mathfrak{X}_k$;
    Add lower-priority cluster to unmatched clusters $\mathfrak{U}$
  end if
end while
return Matched clusters $\mathfrak{M} = \left\{\left(\mathfrak{X}_i, y_i^g\right)\right\}_{i=1}^K$;

3.4. Diffusion Label Matching Refinement

In the SPICE framework, an imperfect feature model can cause similar features to be assigned to truly different clusters, while imperfect cluster heads can result in dissimilar samples being assigned the same cluster label. These issues may eventually lead to the presence of samples from different classes in the same cluster and mismatches between samples and their true labels. Alternatively, these errors can also be caused by the pseudo-label-matching algorithm, which can result in mismatches between the cluster labels and the true labels of the clusters they represent. To overcome these issues, we propose a diffusion model label-matching refinement algorithm to adjust the matching of labels within clusters.

Here, we also use the m-th cluster as an example, similar to our previous selection of $\mathfrak{X}_m^{top}$ and the least confident samples $\mathfrak{X}_m^{btm}$ for m-th cluster being selected as:

$$\mathfrak{X}_m^{btm} = \{(\boldsymbol{x}_i, y_i^s) \mid i \in \text{arglowk}(\boldsymbol{P}_{j,m}, N_{btm}), \boldsymbol{x}_j \in \mathfrak{X}_m\} \tag{13}$$

where the arglowk function $\text{arglowk}\left(\boldsymbol{P}_{j,m}, N_{btm}\right)$ returns the indices of the least N_{btm} confident samples, selected from the indices belonging to the m-th samples in the m-th column of matrix $\boldsymbol{P}_{:,m}$. Similar to $\mathfrak{X}_m^{top}$, zero-shot classification based on true labels for $\mathfrak{X}_m^{btm}$ is also performed using CLIP.

Furthermore, the semantic matching ratios δ_m^{top} and δ_m^{btm} for the top and bottom of the m-th cluster can be represented as:

$$\delta_m^{top} = \frac{1}{N_{top}} \sum_{y \in \mathfrak{X}_m^{top}} \mathbb{1}\left(y = y_i^g\right) \tag{14}$$

$$\delta_m^{btm} = \frac{1}{N_{btm}} \sum_{y \in \mathfrak{X}_m^{btm}} \mathbb{1}\left(y = y_i^g\right) \tag{15}$$

where δ_m^{top} and δ_m^{btm} reflect the matching status of the top and bottom of the m-th cluster. To comprehensively reflect the matching status of the m-th cluster, the overall semantic matching ratio δ_m is defined as:

$$\delta_m = \delta_m^{top} * w_{top} + \delta_m^{btm} * w_{btm} \tag{16}$$

where w_{top} and w_{btm} represent the weights of δ_m^{top} and δ_m^{btm}, respectively, in the overall matching ratio δ_m.

If the overall semantic matching ratio $\delta_m > \lambda_\delta$, where λ_δ is the overall matching threshold, then a high matching degree for the m-th cluster $\mathfrak{X}_m$ implies that the cluster label y_m^g for that is trustworthy. In other cases, further examination is required to determine the matching status of the top and bottom of the m-th cluster. In cases where the matching status of the top δ_m^{top} in the m-th cluster is greater than the top matching threshold λ_{top} but the matching status of the bottom δ_m^{btm} is less than the bottom matching threshold λ_{btm}, it is necessary to evaluate the semantic consistency ratio r_m^{btm} of the least confident samples $\mathfrak{X}_m^{btm}$, which can be defined as:

$$r_m^{btm} = \frac{1}{N_{btm}} \sum_{\boldsymbol{x}_i \in \mathfrak{X}_m^{btm}} \boldsymbol{P}_{i,m} \tag{17}$$

where $\boldsymbol{P}_{i,m}$ denotes the clustering probability of the element located in the i-th row and m-th column of matrix $\boldsymbol{P}$, specifically referring to the probability that $x_i \in \mathfrak{X}_m^{btm}$ belongs to the m-th cluster $\mathfrak{X}_m$. When r_m^{btm} exceeds the confidence threshold σ_{btm}, it implies that even if the matching degree at the bottom level is lower than the threshold, the overall consistency of the m-th cluster $\mathfrak{X}_m$ is sufficiently reliable, thus suggesting that the cluster

label y_m^g as a whole can be trusted. Otherwise, samples $\mathfrak{X}_m^{adj}$ in the m-th cluster with low clustering consistency, denoted as:

$$\mathfrak{X}_m^{adj} = \left\{ \left(\boldsymbol{x}_i, y_i^g \right) \mid \boldsymbol{P}_{i,m} < \sigma_{btm}, \boldsymbol{x}_i \in \mathfrak{X}_m \right\} \tag{18}$$

need to be reassigned to other clusters using CLIP.

If δ_m^{top} is less than the top matching threshold λ_{top}, it suggests a mismatch in the overall clustering. If r_m^{btm} is also lower than the confidence threshold σ_{btm}, we use CLIP to reassign the samples $\mathfrak{X}_m^{adj}$ in the m-th cluster with low clustering consistency to other clusters, indicating that their original cluster labels are no longer valid. Furthermore, when r_m^{btm} exceeds the confidence threshold σ_{btm}, suggesting overall clustering consistency but a mismatch in the overall clustering, we will maintain the original matching cluster labels y_m^g. Following the fine-tuning of each cluster, the resulting fine-tuned clusters, along with their corresponding labels $\mathfrak{M}^* = \left\{ \left(\mathfrak{X}_i^*, y_i^g \right) \right\}_{i=1}^K$ are obtained; $\mathfrak{X}^*$ denotes the cluster that has undergone fine-tuning.

3.5. Synthesis Guided with Matching Labels

For conditional image synthesis, we use a classifier $p_\phi(y \mid \boldsymbol{x})$ to enhance the generator of diffusion models based on clusters $\mathfrak{M}^* = \left\{ \left(\mathfrak{X}_i^*, y_i^g \right) \right\}_{i=1}^K$ with given matching real classes, where $\boldsymbol{x}$ is the input image to the classifier and y is the corresponding output label. The classification network is composed of the feature model $\mathcal{F}$ and the clustering head $\mathcal{C}$ from the clustering stage, along with an additional classification head $\mathcal{C}_{clf}$. As demonstrated in previous works [41,42], a pre-trained diffusion model can be conditioned via the gradients of a classifier. The conditioned reverse denoising process, denoted as $p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ in Equation (4), can be expressed as $p_{\theta,\phi}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}, y)$. In [41,42], the following equation:

$$\begin{aligned} \log p_{\theta,\phi}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}, y) &= \log\left((p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}) p_\phi(y \mid \boldsymbol{x}_t) \right) + B_1 \\ &\approx \log p(z) + B_2, z \sim \mathcal{N}(\mu + \boldsymbol{\Sigma} g, \boldsymbol{\Sigma}) \end{aligned} \tag{19}$$

where $g = \nabla_{\boldsymbol{x}_t} \log p_\phi(y \mid \boldsymbol{x}_t)\big|_{\boldsymbol{x}_t = \mu}$ and $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t) = \boldsymbol{\Sigma}$ for brevity, have been proven. B_1 and B_2 are constants. $p_\phi(y \mid \boldsymbol{x}_t))$ is a shorthand for the classifier $p_\phi(y \mid \boldsymbol{x}_t, t))$ trained on noisy images $\boldsymbol{x}_t$.

4. Results and Discussion

In this section, we evaluate GDUI on different benchmarks in terms of various metrics. First, we describe our benchmark datasets and their evaluation metrics. Second, we describe the baselines and implementation details used in the experiments. Third, we compare GDUI with models that use alternative strategies and show the comparative results. Finally, we conduct ablation experiments and related analyses for the proposed GDUI model.

4.1. Dataset and Evaluation Metrics

We validated the effectiveness of the proposed model on the STL-10 dataset [45]. The STL-10 dataset contains 10 different classes of images, with 500 images per class, totaling 5000 labeled training images and 8000 unlabeled testing images. These images have a resolution of 96 × 96 pixels. During training, we combined and trained the STL-10 dataset's train and test split datasets together. GDUI and its corresponding baselines are quantitatively evaluated for sample quality using the following metrics. We used the Fréchet Inception Distance (FID) [46] as the primary metric to measure the overall sample quality, including both diversity and fidelity, because it is currently the de facto standard metric for evaluating generative models. FID combines the statistical feature differences between the samples generated by the image generation model and the distribution of real images. To ensure more accurate FID calculation, we sampled all 13,000 images from the

STL-10 dataset as samples of real images. To ensure a fair comparison, the distribution of real images for calculating all the metrics was based on the 13,000 images from the STL-10 dataset. In addition, we also used sFID [47], another version of FID that uses spatial features instead of pooled features to capture spatial relationships, as a standard metric for evaluating image quality. For sample fidelity, we used Inception Score (IS) [48] as a supplementary metric, which evaluates the quality of generated images by utilizing a pre-trained Inception-V3 [49] network on ImageNet [50]. The Improved Precision and Recall [51] metrics can separately evaluate the fidelity and diversity of generated samples, where Precision measures the ratio of model samples falling into the data manifold, while Recall measures the ratio of data samples falling into the sample manifold.

4.2. Baselines and Implementation Details

Regarding the choice of baseline, we compare GDUI against both the unconditional diffusion model and the diffusion model guided by a classifier trained on ground truth [5]. When comparing with baselines, we use the same backbone and hyperparameters, ensuring a fair comparison. As the training and sampling steps of diffusion models are typically positively correlated with the sampling results [13], we keep the training and sampling steps consistent at 500 steps when comparing. The details of the initial clustering implementation are as follows. During the representation learning phase of clustering, we employ MoCo-v2 [52] and use ResNet34 [53] as the backbone for feature learning in clustering. Additionally, both the clustering head and classification head consist of two fully connected layers.

4.3. Comparisons

On one hand, we compare GDUI to an unconditional diffusion model and a diffusion model trained on datasets with ground-truth labels. On the other hand, we also compare GDUI to other state-of-the-art methods. Due to the potential impact of different compute budgets on fair comparisons, we utilize the same compute budget for all compared models. The comparison results on the STL-10 dataset are shown in the Table 1. We observe significant improvements of GDUI over unconditional diffusion model in evaluation metrics, such as 37.9% increase in FID score, 19.3% increase in IS score, and 25.4% increase in precision. More notably, GDUI outperforms diffusion model trained on ground-truth labels in all evaluation metrics except for the Recall score, where it performs slightly worse. Furthermore, GDUI surpasses the comparison models in all evaluation metrics except for the Recall score. We believe that GDUI achieves this effect by aligning images with similar features better with label information. We explicitly evaluate changes in sample quality under different training iterations as shown in Figure 2. The results indicate that GDUI outperforms the unconditional diffusion model on all evaluation metrics except for Recall, with a significant margin observed in most cases.

Table 1. Comparison of GDUI on the whole merged STL-10 dataset, combining the train and test split datasets. The best results are highlighted in **bold**.

Diffusion Method	FID↓	sFID↓	IS↑	Precision↑	Recall↑
Unconditional	26.81	48.74	9.76	0.59	**0.45**
Ground-truth guidance	17.20	48.55	11.40	0.72	0.41
IGGAN [54]	21.39	49.53	10.71	0.65	0.39
DDGAN [55]	21.79	49.12	10.34	0.69	0.40
TransGAN [56]	18.28	50.30	11.34	0.64	0.42
GDUI	**16.66**	**48.44**	**11.64**	**0.74**	0.39

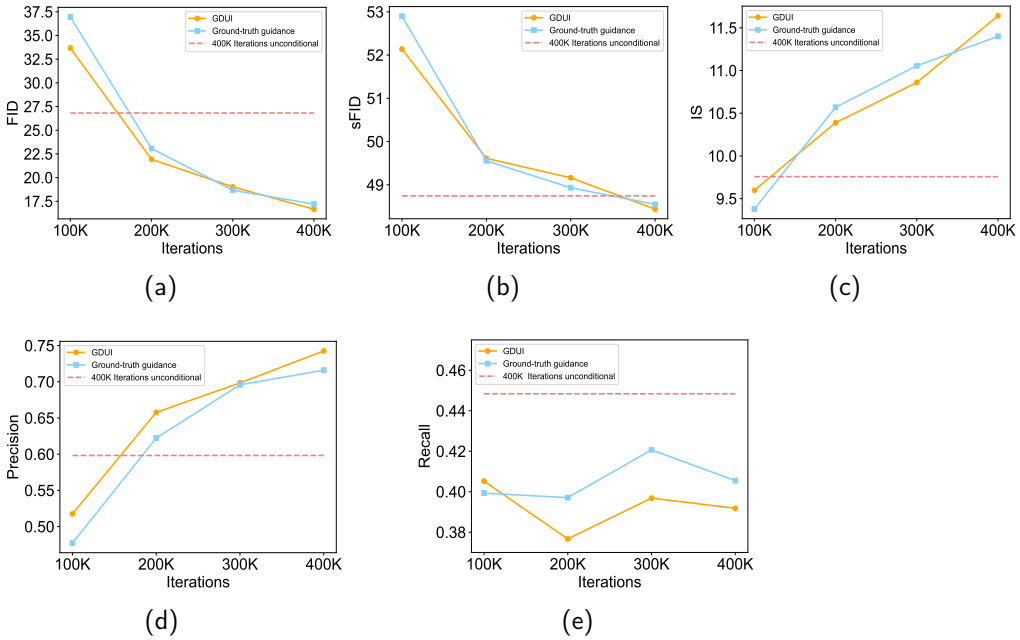


Figure 2. Changes in sample quality under different iterations. Throughout the entire training process, GDUI outperforms the diffusion model trained with ground-truth labels in terms of FID, sFID, IS, and precision metrics in many cases, indicating that the superior performance of GDUI is not just a coincidence. Moreover, both GDUI and the diffusion model trained with ground-truth labels outperform the unconditioned diffusion model trained for 400K iterations, which achieved the best performance throughout the entire training process, in terms of FID, sFID, IS, and precision metrics. (**a**) FID at different iterations. (**b**) sFID at different iterations. (**c**) IS at different iterations. (**d**) Precision at different iterations. (**e**) Recall at different iterations.

Compared to the diffusion model based on ground-truth labels, GDUI maintains an advantage in FID, IS, sFID, and precision metrics on most training stages, especially as the number of iterations increases. However, GDUI still lags behind the diffusion model based on ground-truth labels by 2% in Recall.

Qualitative results on STL-10 are shown in Figure 3. It can be observed that the perceptual quality of GDUI-generated samples is higher than those generated by diffusion models based on unconditional and ground-truth labels. For example, GDUI generates images with relatively clear structures of cars and better shapes of horses compared to other diffusion models.

4.4. Ablation Study

In this section, we conduct experiments to investigate the effects of different parameters and components on the GDUI framework.

4.4.1. Component Ablation

To guide the diffusion models to generate high-quality images on unlabeled datasets, we utilize three key methods: pseudo-label generation (PLG), pseudo-label matching (PLM), and label-matching refinement (LMR). So, we perform four gradually changing settings to validate the effectiveness of the three components: (1) We use unconditional diffusion models as a baseline for ablation experiments, as described in Section 4.2. (2) We utilize the pseudo-label generation (PLG) method to generate pseudo-labels for unlabeled datasets, and directly train and generate images on them using the guided diffusion model. (3) To generate images of specific real categories, we utilize the pseudo-label-matching (PLM) method to train and generate images. (4) Finally, we employ the label-matching

refinement (LMR) method to further improve the performance by adjusting irrelevant semantic information of the images.

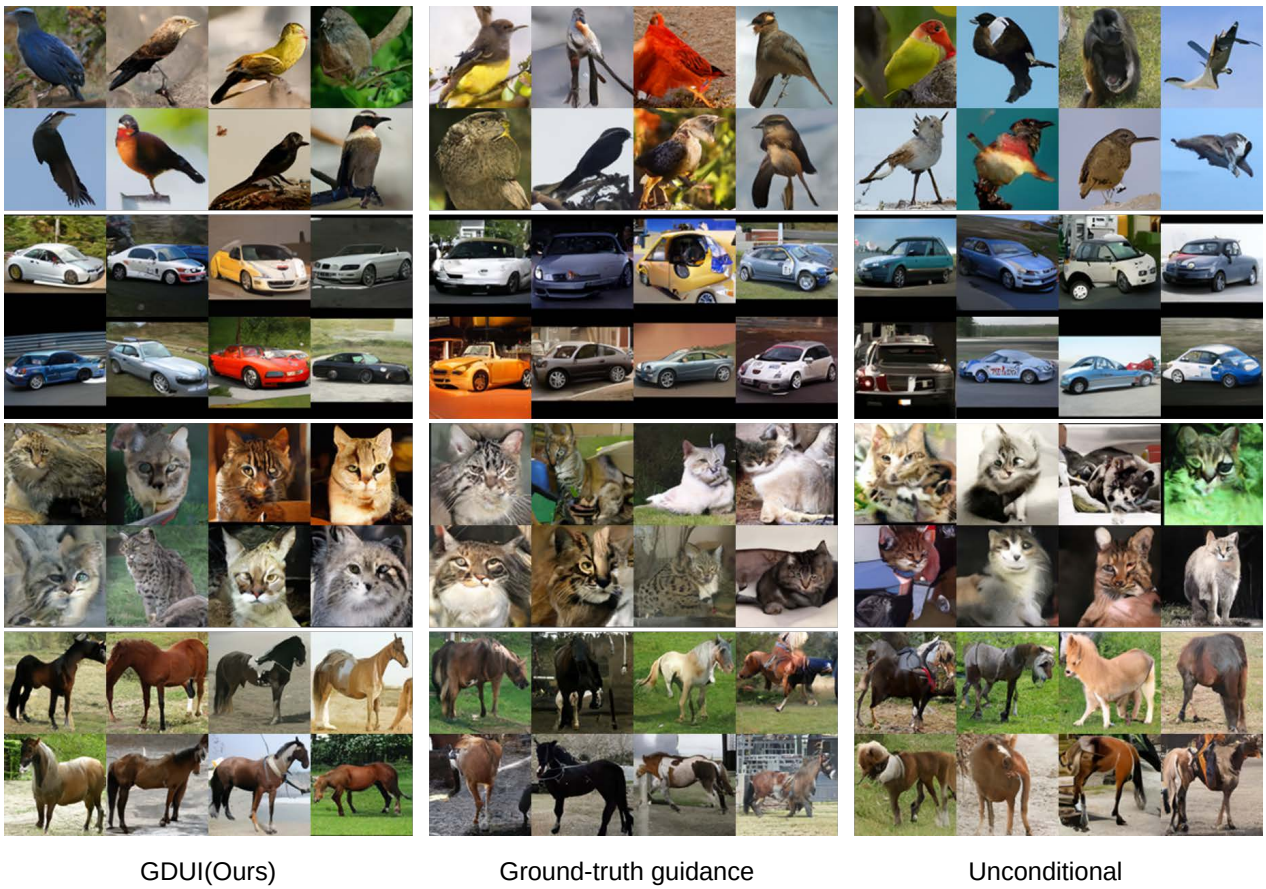


Figure 3. Qualitative comparisons on STL-10 dataset. Samples from our diffusion model GDUI (FID 16.66, **left**) compared to samples from a diffusion model based on ground-truth labels (FID 17.20, **middle**) and samples from a diffusion model based on an unlabeled dataset (FID 26.81, **right**).

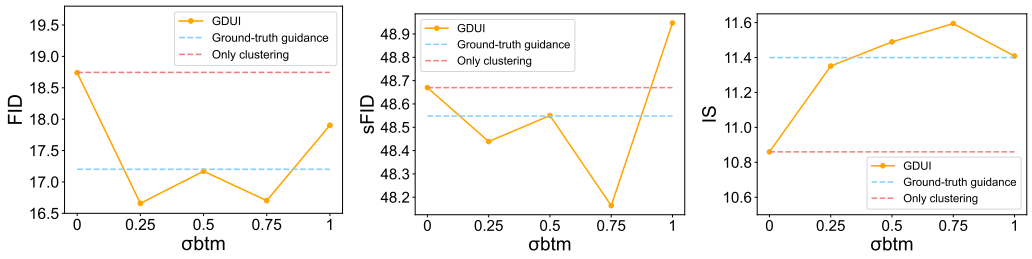
We show the results in Table 2. The initial performance of the baseline, i.e., the unconditional diffusion model, on all metrics except Recall is relatively low. Moreover, since it is built on an unlabeled dataset, it cannot generate images of specific real categories. By adopting the pseudo-label generation (PLG) method, the quality of the generated images improves in terms of FID, sFID, IS, and precision metrics. Particularly, there is a significant improvement in FID and precision metrics. However, the problem of generating images of specific real categories is not completely solved since only pseudo-labels are generated. The aforementioned problem is addressed by utilizing the pseudo-label-matching (PLM) method. Meanwhile, since the pseudo-label-matching (PLM) method only matches the clusters with pseudo-labels to the actual classes, without better aligning the images and labels, the quality of the generated images does not improve. Finally, we utilize the label-matching refinement (LMR) method to adjust the semantic information within the clusters with actual class information, in order to better align the images with the semantic information. It further improves the image quality while maintaining the ability to generate images of specific real categories.

Table 2. Ablation studies of GDUI on the whole STL10 dataset. Here, baseline refers to the unconditional diffusion model trained for 400K iterations on the unlabeled dataset. Qualitative comparisons are also conducted to evaluate whether the models can generate images of specific real categories. The best results are highlighted in **bold**.

Method	Real Class	FID ↓	sFID↓	IS ↑	Precision↑	Recall↑
Baseline	NO	26.81	48.74	9.76	0.59	**0.45**
+PLG	NO	18.74	48.67	10.86	0.69	0.38
+PLM	**YES**	18.73	48.66	10.85	0.69	0.38
+LMR	**YES**	**16.66**	**48.44**	**11.64**	**0.74**	0.39

4.4.2. Effect of Confidence Threshold

The confidence threshold σ_{btm} is an important hyperparameter of the label-matching refinement method, which determines whether clustering needs to be fine-tuned and the range involved in the fine-tuning process. Thus, we investigate how the confidence threshold σ_{btm} influences the quality of images generated by GDUI, and analyze the results accordingly. As depicted in Figure 4, for the majority of cases where the confidence threshold σ_{btm} ranges from 0 to 1, GDUI exhibits superior performance over the unconditional diffusion model across various evaluation metrics. It is worth noting that in the majority of cases, GDUI also exhibits superior performance over the diffusion model based on ground-truth datasets in all evaluation metrics except Recall. As can also be observed from Figure 4, when the confidence threshold σ_{btm} approaches a small value close to 0, the overall performance of GDUI degrades to be similar to only using the pseudo-label generation (PLG) method. This phenomenon can be understood as the label-matching refinement method determining that no fine-tuning is needed for the labels within a cluster when the confidence threshold σ_{btm} is set to a low value. When σ_{btm} tends to 1 with a large value, the overall performance of GDUI becomes similar to the diffusion model based on ground-truth datasets. This can be interpreted as the label-matching refinement method considering that almost all image labels need to be fine-tuned.



(a) FID at various σ_{btm}. (b) sFID at various σ_{btm}. (c) IS at various σ_{btm}.

(d) Precision at various σ_{btm}. (e) Recall at various σ_{btm}.

Figure 4. Ablation results of various σ_{btm} values on the STL10 dataset. As σ_{btm} increases, the general trend for all metrics is to first improve and then worsen.

5. Limitations and Future Work

In this study, the GDUI significantly improves the fidelity and diversity of generated images on target unlabeled datasets, and even outperforms the diffusion model based on ground-truth labels in most cases. However, both the clustering method we used and existing clustering methods assume prior knowledge of the number of clusters K, and in order to match with the ground-truth labels, we also need to know the names of the true classes, which may not always be available in some cases. In the future, our method could be combined with image-to-text approaches to obtain the true textual information, not just the class information, for unlabeled datasets. Furthermore, our method can also be applied to tasks such as graph property prediction [57] and improving the adversarial robustness in classification tasks [58]. In future personalized photo customization [59], our approach is poised to produce higher-quality photos more effectively, while also better aligning with user preferences and specific needs. It provides assistance for the issue of annotating medical images, which requires domain experts [60,61]. In addition, the diffusion model in this paper is based on a pixel-wise diffusion model, which incurs high training costs. However, the GDUI framework is a generic diffusion model framework designed for unlabeled datasets, and it can be adapted to be compatible with latent-based diffusion models in the future to achieve the goal of reducing training costs.

6. Conclusions

We have presented GDUI, a generic guided diffusion model framework designed for unlabeled datasets. The framework consists of three stages: encoding images into semantically meaningful latent vectors and mapping semantic information, filtering out irrelevant information, and generating images based on matched real semantic information. Our experiments have demonstrated that compared to the diffusion model based on unlabeled datasets, GDUI achieved a 37.9% improvement in FID, a 19.3% improvement in IS, and a 25.4% improvement in precision. This indicates a significant enhancement in the quality of generated samples. Additionally, the experiments demonstrate that GDUI can generate high-quality images of specified categories on unlabeled datasets, and even outperforms the diffusion model based on ground-truth labels. Our future goal is to achieve higher-quality image generation on large-scale and multi-category unlabeled datasets.

Author Contributions: Conceptualization, X.X.; Formal analysis, X.X.; Funding acquisition, J.Z.; Investigation, X.X.; Methodology, X.X.; Project administration, J.Z.; Resources, J.Z.; Software, X.X.; Supervision, J.Z.; Validation, X.X.; Visualization, X.X.; Writing—original draft, X.X.; Writing—review and editing, X.X. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China under Grants 62006131 and 62071260, the National Natural Science Foundation of Zhejiang Province under Grants LQ21F020009 and LZ22F020001.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Po, R.; Yifan, W.; Golyanik, V.; Aberman, K.; Barron, J.T.; Bermano, A.H.; Chan, E.R.; Dekel, T.; Holynski, A.; Kanazawa, A.; et al. State of the art on diffusion models for visual computing. *arXiv* **2023**, arXiv:2310.07204.
2. Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; Guo, B. Vector quantized diffusion model for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10696–10706.
3. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv* **2022**, arXiv:2204.06125.
4. Nichol, A.Q.; Dhariwal, P. Improved denoising diffusion probabilistic models. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 8162–8171.
5. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.

6. Ho, J.; Salimans, T. Classifier-Free Diffusion Guidance. In Proceedings of the NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications, Virtual Event, 14 December 2021.
7. Mall, U.; Hariharan, B.; Bala, K. Change event dataset for discovery from spatio-temporal remote sensing imagery. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27484–27496.
8. Shin, H.; Kim, H.; Kim, S.; Jun, Y.; Eo, T.; Hwang, D. SDC-UDA: Volumetric Unsupervised Domain Adaptation Framework for Slice-Direction Continuous Cross-Modality Medical Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7412–7421.
9. Bordes, F.; Balestriero, R.; Vincent, P. High Fidelity Visualization of What Your Self-Supervised Representation Knows About. In *Transactions on Machine Learning Research*; OpenReview.net: Amherst, MA, USA, 2022.
10. Hu, V.T.; Zhang, D.W.; Asano, Y.M.; Burghouts, G.J.; Snoek, C.G. Self-guided diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18413–18422.
11. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual Event, 18–24 July 2021; pp. 8748–8763.
12. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
13. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 36479–36494.
14. Wang, Z.; Zhang, Z.; Zhang, X.; Zheng, H.; Zhou, M.; Zhang, Y.; Wang, Y. DR2: Diffusion-based Robust Degradation Remover for Blind Face Restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1704–1713.
15. Li, Y.; Fan, Y.; Xiang, X.; Demandolx, D.; Ranjan, R.; Timofte, R.; Van Gool, L. Efficient and explicit modelling of image hierarchies for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18278–18289.
16. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. In Proceedings of the International Conference on Machine Learning, PMLR, New Orleans, LA, USA, 18–24 June 2021; pp. 8821–8831.
17. Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 22500–22510.
18. Wang, W.; Bao, J.; Zhou, W.; Chen, D.; Chen, D.; Yuan, L.; Li, H. Semantic image synthesis via diffusion models. *arXiv* **2022**, arXiv:2207.00050.
19. Han, X.; Zheng, H.; Zhou, M. Card: Classification and regression diffusion models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 18100–18115.
20. Kim, G.; Kwon, T.; Ye, J.C. Diffusionclip: Text-guided diffusion models for robust image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 2426–2435.
21. Nichol, A.Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; Chen, M. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 16784–16804.
22. Sheynin, S.; Ashual, O.; Polyak, A.; Singer, U.; Gafni, O.; Nachmani, E.; Taigman, Y. kNN-Diffusion: Image Generation via Large-Scale Retrieval. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
23. Blattmann, A.; Rombach, R.; Oktay, K.; Müller, J.; Ommer, B. Retrieval-augmented diffusion models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 15309–15324.
24. Zhou, Y.; Zhang, R.; Chen, C.; Li, C.; Tensmeyer, C.; Yu, T.; Gu, J.; Xu, J.; Sun, T. Towards language-free training for text-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 17907–17917.
25. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.A.; Bottou, L. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
26. Ji, P.; Zhang, T.; Li, H.; Salzmann, M.; Reid, I. Deep subspace clustering networks. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
27. Tian, K.; Zhou, S.; Guan, J. Deepcluster: A general clustering framework based on deep learning. In Proceedings of the Machine Learning and Knowledge Discovery in Databases, Skopje, Macedonia, 18–22 September 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 809–825.
28. Jiang, Z.; Zheng, Y.; Tan, H.; Tang, B.; Zhou, H. Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 19–25 August 2017; pp. 1965–1972.
29. Zhou, P.; Hou, Y.; Feng, J. Deep adversarial subspace clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1596–1604.

30. Zhang, J.; Li, C.G.; You, C.; Qi, X.; Zhang, H.; Guo, J.; Lin, Z. Self-supervised convolutional subspace clustering network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5473–5482.
31. Niu, C.; Shan, H.; Wang, G. Spice: Semantic pseudo-labeling for image clustering. *IEEE Trans. Image Process.* **2022**, *31*, 7264–7278. [CrossRef]
32. Han, S.; Park, S.; Park, S.; Kim, S.; Cha, M. Mitigating embedding and class assignment mismatch in unsupervised image classification. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 768–784.
33. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
34. Banterle, F.; Marnerides, D.; Bashford-Rogers, T.; Debattista, K. Self-Supervised High Dynamic Range Imaging: What Can Be Learned from a Single 8-bit Video? *ACM Trans. Graph.* **2024**, *just accepted*. [CrossRef]
35. Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; Gao, J. Vinvl: Revisiting visual representations in vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5579–5588.
36. Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; Hoi, S.C.H. Align before fuse: Vision and language representation learning with momentum distillation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9694–9705.
37. Huang, Y.; Wang, Y.; Zeng, Y.; Wang, L. MACK: Multimodal aligned conceptual knowledge for unpaired image-text matching. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 7892–7904.
38. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual Event, 18–24 July 2021; pp. 4904–4916.
39. Ao, T.; Zhang, Z.; Liu, L. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. *ACM Trans. Graph.* **2023**, *42*. [CrossRef]
40. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
41. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 6–11 July 2015; pp. 2256–2265.
42. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
43. Caron, M.; Bojanowski, P.; Joulin, A.; Douze, M. Deep clustering for unsupervised learning of visual features. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 132–149.
44. Qi, D.; Chen, S.; Sun, X.; Luan, R.; Tong, D. A multiscale convolutional gragh network using only structural information for entity alignment. *Appl. Intell.* **2023**, *53*, 7455–7465. [CrossRef]
45. Coates, A.; Ng, A.; Lee, H. An analysis of single-layer networks in unsupervised feature learning. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 215–223.
46. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6626–6637.
47. Nash, C.; Menick, J.; Dieleman, S.; Battaglia, P. Generating images with sparse representations. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual Event, 18–24 July 2021; pp. 7958–7968.
48. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. *Adv. Neural Inf. Process. Syst.* **2016**, *29*.
49. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
50. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Toulouse, France, 2009; pp. 248–255.
51. Kynkäänniemi, T.; Karras, T.; Laine, S.; Lehtinen, J.; Aila, T. Improved precision and recall metric for assessing generative models. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
52. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved baselines with momentum contrastive learning. *arXiv* **2020**, arXiv:2003.04297.
53. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
54. Zhang, Z.; Hua, Y.; Wang, H.; McLoone, S. Improving the Fairness of the Min-Max Game in GANs Training. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2024; pp. 2910–2919.
55. Xiao, Z.; Kreis, K.; Vahdat, A. Tackling the Generative Learning Trilemma with Denoising Diffusion GANs. In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2022.
56. Jiang, Y.; Chang, S.; Wang, Z. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 14745–14758.

57. Liu, G.; Inae, E.; Zhao, T.; Xu, J.; Luo, T.; Jiang, M. Data-Centric Learning from Unlabeled Graphs with Diffusion Model. In *Advances in Neural Information Processing Systems*; Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: Nice, France, 2023; Volume 36, pp. 21039–21057.
58. Ouyang, Y.; Xie, L.; Cheng, G. Improving Adversarial Robustness Through the Contrastive-Guided Diffusion Process. In Proceedings of the 40th International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J., Eds.; JMLR: Cambridge, MA, USA, 2023; Volume 202, pp. 26699–26723.
59. Zhang, Y.; Dong, W.; Tang, F.; Huang, N.; Huang, H.; Ma, C.; Lee, T.Y.; Deussen, O.; Xu, C. ProSpect: Prompt Spectrum for Attribute-Aware Personalization of Diffusion Models. *ACM Trans. Graph.* **2023**, *42*. [CrossRef]
60. Li, Y.; Shao, H.C.; Liang, X.; Chen, L.; Li, R.; Jiang, S.; Wang, J.; Zhang, Y. Zero-Shot Medical Image Translation via Frequency-Guided Diffusion Models. *IEEE Trans. Med. Imaging* **2024**, *43*, 980–993. [CrossRef] [PubMed]
61. He, Y.; Ge, R.; Qi, X.; Chen, Y.; Wu, J.; Coatrieux, J.L.; Yang, G.; Li, S. Learning Better Registration to Learn Better Few-Shot Medical Image Segmentation: Authenticity, Diversity, and Robustness. *IEEE Trans. Neural Networks Learn. Syst.* **2024**, *35*, 2588–2601. [CrossRef]