



Title	A study on latent representation learning for understanding, generation, and copyright protection of artificially created works
Author(s)	蘭, 子文
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第16997号
Issue Date	2026-03-25
DOI	https://doi.org/10.14943/doctoral.k16997
Doc URL	https://hdl.handle.net/2115/99841
Type	doctoral thesis
File Information	Ziwen_Lan.pdf, 全文



A thesis for the degree of Doctor of Philosophy

**A Study on Latent Representation Learning for
Understanding, Generation, and Copyright
Protection of Artificially Created Works**

人工創作物の理解・生成および著作権保護のための
潜在表現学習に関する研究



Ziwen Lan

Graduate School of Information Science and Technology

Hokkaido University

March, 2026

Contents

1	Introduction	10
1.1	Background	10
1.2	Proposition in This Thesis	12
1.3	Organization of This Thesis	13
1.4	Licensing Information for Artificially Created Works in This Thesis	14
2	Related Works	15
2.1	Introduction	15
2.2	Semantic Understanding Methods for Artificially Created Works	15
2.3	Personalized Text-to-image Diffusion Models	17
2.4	Copyright Protection Methods for Artificially Created Works	20
2.5	Problems to Be Solved in This Thesis	21
2.6	Conclusion	22
3	Latent Representation-based Semantic Understanding Methods for Artificially Created Works	24
3.1	Latent Representation-based Semantic Understanding Method for Real-World Images Using Image Captioning	27
3.1.1	Introduction	27
3.1.2	Proposed Method	28
3.1.2.1	ADD-GCN	28
3.1.2.2	Import of Captioning Features	31
3.1.2.3	Final Classification and Loss	31
3.1.3	Experiments	31
3.1.4	Conclusions	32
3.2	Latent Representation-based Semantic Understanding Method for Artificially Created Works Using Image Captioning	34
3.2.1	Introduction	34
3.2.2	Proposed Method	38

3.2.2.1	New Captioning Framework for Anime Illustration	38
3.2.2.2	GCN-based Multi-modal Multi-label Attribute Classification Using Domain-specific Semantic Features	39
3.2.2.3	Final Classification and Loss	40
3.2.3	Experiments	41
3.2.3.1	Training of Captioning Model for Anime Illustrations	41
3.2.3.2	Training of Whole Multi-label Classification Model	41
3.2.3.3	Comparison Methods and Evaluation Metrics	41
3.2.3.4	Experimental Results	43
3.2.4	Conclusions	44
3.3	Latent Representation-based Semantic Understanding Method for Artificially Created Works Considering Label Hierarchy	45
3.3.1	Introduction	45
3.3.2	Proposed Method	48
3.3.2.1	New Captioning Framework for Anime Illustrations	50
3.3.2.2	GCN-based Multi-label Attribute Classification	50
3.3.2.3	Final Classification and Loss	52
3.3.3	Experimental Results	53
3.3.3.1	Training Settings of Captioning Model for Anime Illustrations	53
3.3.3.2	Anime Illustration Dataset and Construction of Label Hierarchy	54
3.3.3.3	Experimental Conditions, Comparison Methods, and Evaluation Metrics	55
3.3.3.4	Experimental Results and Discussions	57
3.3.4	Conclusion	61
4	Adversarial Perturbation Based Continual Personalization Diffusion Model for Mitigating Concept Bleed-through	62
4.1	Introduction	62
4.2	Methodology	67
4.2.1	Parameter Modification in Continual LoRA	67
4.2.2	Self-regularization	68
4.2.3	Training on Images with Adversarial Perturbations	69
4.3	Experiments	70
4.3.1	Main Experimental Settings	70
4.3.2	Experimental Results	74
4.3.3	Additional Experiments	76

4.3.4	Comparison with Selective Forgetting Methods	77
4.4	Conclusion	78
5	Latent Representation-based Copyright Protection Methods for Artificially Created Works Using Adversarial Perturbation	79
5.1	Introduction	79
5.2	Preliminary	84
5.2.1	Threat Model	84
5.2.2	Diffusion Process	84
5.2.3	Prompt-guided Sampling in Diffusion Models	85
5.3	Methodology	85
5.3.1	Adversarial Attack for Diffusion Models	87
5.3.2	Adversarial Attack Manipulating Semantic Information	89
5.3.3	Final Loss	90
5.4	Experiments	90
5.4.1	Experimental Settings	90
5.4.2	Experimental Results	94
5.4.3	Effectiveness Evaluation	94
5.5	Limitations	103
5.6	Conclusions	103
6	Conclusions	105
6.1	Overview of the Proposition in This Thesis	105
6.2	Future Directions	106
	Acknowledgements	108
	Bibliography	109
	Achievements of the Author	117

List of Figures

2.1	Research map of related works and position of each chapter in this thesis.	23
3.1.1	Overview of our model. The captioning features from the pre-trained captioning model are processed by the pooling layer and then combined with the output H of the static GCN network in the ADD-GCN model.	30
3.1.2	An example of ADD-GCN and our method. All the confidence scores are sorted in descending order. Labels in red font are the ground truth labels, and labels in the red box represent the final classification.	33
3.2.1	Overview of the GCN-based multi-modal multi-label attribute classification model for anime illustrations using domain-specific semantic features. We train a captioning model for anime illustration by using the captioning of real-world images and corresponding animated style images generated by the image style transformation model. We employ this captioning model to obtain the semantic feature v_{cap} of anime illustration. The semantic feature v_{cap} is used in the GCN-based multi-label classification model to improve the classification accuracy. Source: Anonymous authors, https://pixabay.com . Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). Modified by the author.	37
3.2.2	An example of the ablation experiment. All the scores s are sorted in descending order. We mark the ground truth labels in red font, and the final classification results of each method in the red box. Source: Anonymous authors, https://pixabay.com . Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). Modified by the author.	43

3.3.1 Overview of the proposed model for hierarchical multi-label attribute classification of anime illustrations. Source: Anonymous authors, https://pixabay.com . Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). Modified by the author.	49
3.3.2 Part of the label hierarchy of anime illustration dataset [1].	54
3.3.3 Examples of the effect of the captioning model on the final classification in the ablation experiments. The captioning results of each illustration generated by the captioning models used in the proposed method and CM2 are shown in the boxes of blue and orange backgrounds, respectively. We also show the predicted scores of the labels using histograms. Labels are considered positive if the predicted score is greater than 0.5. The horizontal axis labels of the histograms are the concatenated sets of the top-7 highest scoring labels predicted by the proposed method and CM2, and the ground truth labels are marked in red font. Source: Anonymous authors, https://pixabay.com . Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). Modified by the author.	59
3.3.4 An example of the effect of the hierarchical structure on the final classification in the ablation experiments. Source: Anonymous authors, https://pixabay.com . Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). Modified by the author.	60
4.1 Examples of the concept bleed-through problem in the continual learning of personalized generative models. When handling nearby concepts (e.g., <i>puppy</i> and <i>husky</i>) or related concepts (e.g., <i>dog</i> and <i>petshop</i>), although the prompt only involves a single learned concept, the generated image is clearly influenced by the other learned concepts.	64
4.2 Overview of the proposed method. We update the key-value (K - V) projection in the U-Net cross-attention modules of stable diffusion using a continual, self-regulating low-rank weight adaptation. In addition, we add adversarial perturbations to the training data to achieve effective concept separation in the latent space.	66

4.3	Qualitative results of experiments. All compared methods were affected by concept bleed-through in the learning sequence when dealing with both the nearby concepts (e.g., table and chair) and the related concepts (e.g., vase and flower). In contrast, our proposed method ensured accurate representation of the concepts in the image.	71
4.4	Examples of the additional experiment with different numbers of concepts in the training images. We trained the model with ten concepts in the learning sequence. We selected STAMINA [2] as the baseline method, and after each new concept was learned, the model was used to generate images of the most initially learned concept (<i>vase</i>).	73
4.5	Examples of generated images using different target image categories for perturbation. Semantically dissimilar pairs lead to better concept separation.	75
5.1	Examples of using the personalization method of the text-to-image diffusion model to generate images. Previous adversarial attack methods [3–6] introduce imperceptible perturbations to input images, causing personalized diffusion models trained on these images to generate outputs with significant noise or visible watermark. However, the underlying semantic information is still preserved, allowing simple image processing techniques to restore high-quality images. In contrast, the objective of our method, SeMTAD, is degrade the quality of the generated images but also to manipulate them such that the semantic information differs significantly from that of the input images . (In the figure, G represents the perturbation generation)	81
5.2	Our attack scenario. Users who collect training images from the Web may unknowingly include adversarial examples uploaded by the attacker. When this occurs, the personalized model will generate low-quality images with incorrect content, as specified by the attacker.	83
5.3	Our attack scenario. Users collecting images for training from the Web may unknowingly include adversarial examples uploaded by attackers. In this scenario, the personalized model will generate low-quality images with incorrect content, as dicdted by the attacker.	86

5.4	Attack results against DreamBooth [7]. We compared the images generated by SeMTAD and the benchmark methods on two datasets, LSUN-cat [8] and UECFoodPix-croissant [9]. SeMTAD successfully altered the content of the images, transforming a cat into an airplane and a croissant into person, respectively.	96
5.5	Attack results against Textual Inversion [10]. We compared the images generated by SeMTAD and the benchmark methods on two datasets, LSUN-cat [8] and UECFoodPix-croissant [9]. SeMTAD successfully altered the content of the images, transforming a cat into an airplane and a croissant into person, respectively.	97
5.6	Results of ablation studies on LUSN-cat [8] with respect to changing the number of adversarial examples included in training data.	98
5.7	Example of the ablation study, illustrating the impact of varying the number of adversarial examples in the training data. Green boxes represent clean images, while red boxes indicate adversarial examples included in the training data. The attack target is DreamBooth [7].	99
5.8	Examples of ablation studies on category pair cat→dog and car→airplane. All training images are adversarial examples. The attack target is DreamBooth [7].	100
5.9	Results of the robustness validation on category pairs croissant→person. We applied JPEG compression [11], TVM [12], and SR [13] as defense methods on our adversarial examples. The attack target is DreamBooth [7].	104

List of Tables

3.1.1 Comparison of our model and other models. The best results are marked as bold.	32
3.2.1 Performance comparison between our model and other image classification models. We mark the best results as bold.	42
3.2.2 The ablation experimental results. We can see the influence of the accuracy of the captioning model (BLEU) on the final classification (mAP). In the ablation experiment, we randomly selected 100 anime illustrations with ground truth captionings from Danbooru2020 dataset and calculated BLEU score and mAP. . . .	42
3.3.1 Performance comparison between our model and other image classification models. We mark the best results as bold.	56
3.3.2 Ablation experimental results. In the ablation experiment, we randomly selected 100 anime illustrations with ground truth captions from the Danbooru 2020 dataset and calculated BLEU score [14] and mAP.	57
4.1 Comparison of MMD and CLIP scores across different methods.	72
4.2 Performance of our method on different target images for the perturbation optimization of <i>chair</i> . “-” means we do not adopt any perturbation to the training images.	74
4.3 Comparison between selective forgetting methods and ours method.	76
5.1 Experimental results for DreamBooth [7].	93
5.2 Experimental results for Textual Inversion [10].	93
5.3 Results of ablation studies on different category pairs. We set the number of adversarial examples in the training data to 5. The attack target is DreamBooth [7].	94

5.4	Results of the ablation studies on different category pairs. We set the number of adversarial examples included in the training data to 5. The target of the attack is Textual Inversion [10].	95
5.5	Effects of our attack under different perturbation budgets. The target of the attack is DreamBooth [7].	101
5.6	Effects of our attack under different perturbation budgets. The target of the attack is Textual Inversion [10].	101
5.7	Results of ablation studies on ImageNet-cat [15] with and without incorporating semantic segmentation (SAM2) into SeMTAD. The attack target is DreamBooth [7].	102
5.8	Experimental results of the validation of the image fidelity.	104

Chapter 1

Introduction

The background and purpose of this thesis and the organization of this thesis are presented in this chapter.

1.1 Background

In recent years, deep learning has achieved remarkable progress in various domains such as computer vision, natural language processing, and multimodal content understanding. Large-scale models, including ResNet [16], BERT [14], ViT [17], CLIP [18], Stable Diffusion [19], and GPT-based systems [20], have demonstrated outstanding performance through large-scale data-driven training paradigms. These models inherently rely on latent representations that encode semantic structures underlying images, text, and multimodal signals. As AI technologies expand into creative fields, latent representation learning has become increasingly important for understanding, generating, and protecting artificially created works.

In this thesis, *artificially created works* refer to two broad categories: (1) human-created works, such as anime illustrations, drawings, and other manually crafted content, and (2) AI-generated works, produced by modern generative models such as diffusion models. Although these categories differ in their creation process, both share the essential characteristic of being governed by latent semantic structures that encode attributes, styles, conceptual dependencies, and hierarchical relations. As generative AI becomes increasingly pervasive, understanding and controlling these latent structures have emerged as critical technical and societal challenges.

The rapid deployment of generative systems has revealed limitations in the semantic under-

standing of artificially created works. For human-created content, semantic attributes often exhibit complex multi-label and hierarchical dependencies. Unlike natural images, artificially created works intentionally contain stylized, abstract, or exaggerated features. These characteristics complicate automated semantic understanding, as conventional recognition models struggle to capture fine-grained semantics, attribute interactions, and domain-specific expression rules [16,21].

At the same time, AI-generated works, particularly those produced by personalized diffusion models [7, 10, 22], have become widely used in creative and industrial applications. Users increasingly interact with generative systems that continually incorporate personalized concepts, such as original characters or custom styles [2,23,24]. Although personalization enhances usability, continual learning introduces a serious technical problem known as *concept bleed-through*, in which newly added concepts overwrite or interfere with previously learned ones. This instability undermines the reliability of generative systems and highlights their limited capability in maintaining robust and disentangled latent structures.

In parallel, concerns regarding copyright protection have intensified. The increasing use of web-scraped datasets containing copyrighted artworks has raised ethical and legal concerns about unauthorized training and stylistic imitation. As generative models refine their ability to learn and reproduce latent representations, protecting human-created works from misuse has become essential. Ensuring content safety requires methods that can prevent generative models from extracting or learning sensitive latent information, while preserving the visual fidelity of protected content.

These developments reveal a shared bottleneck across semantic understanding, generative controllability, and protection: current methods are limited in their ability to appropriately learn, regulate, and defend latent semantic representations of artificially created works. Motivated by these issues, this thesis places latent representation learning at its core and aims to address the challenges of semantic modeling, robust generation, and copyright protection within a unified framework.

1.2 Proposition in This Thesis

The purpose of this thesis is to develop a unified research framework based on latent representation learning to address three fundamental challenges associated with artificially created works: Improving semantic understanding performance, ensuring generative controllability and robustness, and enhancing content safety and copyright protection. These three challenges represent the essential capabilities required for next-generation intelligent systems dealing with artificially created works.

First, this thesis investigates latent representation–based semantic modeling for human-created works such as anime illustrations. To capture complex attribute relations and semantic dependencies, this thesis employs hierarchical modeling, graph-based representations, and multimodal fusion techniques. These approaches enable the extraction of structured latent features that represent fine-grained semantic relationships.

Second, this thesis explores generative controllability in personalized diffusion models by addressing the concept bleed-through problem. Existing personalization methods often modify model parameters or architectures but fail to maintain separability in latent spaces. To overcome this limitation, this thesis proposes an adversarial perturbation strategy that manipulates training images at the data level. This ensures clear separation between new and previously learned concepts without adding model complexity, leading to improved generative stability.

Third, this thesis proposes a latent representation–oriented protection framework to defend human-created works against unauthorized training. By introducing small but semantically effective perturbations that alter latent representations while preserving visual fidelity, the proposed method prevents generative models from extracting stylistic or conceptual information. This provides practical and effective means of copyright protection in the era of large-scale generative AI.

The contributions of this thesis can be summarized as follows:

- This thesis presents latent representation learning methods for modeling semantic structures in artificially created works, enabling improved understanding of hierarchical and multimodal attributes.
- This thesis proposes an adversarial perturbation–based continual personalization method for diffusion models, effectively mitigating concept bleed-through and enhancing genera-

tive robustness.

- This thesis introduces a latent-space-oriented protection technique that prevents unauthorized training of generative models while maintaining the perceptual quality of human-created works.

1.3 Organization of This Thesis

The remainder of this thesis is organized as follows.

In Chapter 2, related works on latent representation learning, semantic modeling, generative controllability, and protection techniques are introduced. The research problems addressed in this thesis are clarified.

In Chapter 3, latent representation-based semantic understanding methods for human-created works are presented. This chapter focuses on image captioning, hierarchical attribute modeling, multi-label classification, and multimodal fusion for extracting structured semantic features.

In Chapter 4, a method for improving generative controllability in personalized diffusion models is presented. An adversarial perturbation-based continual personalization strategy is proposed to mitigate concept bleed-through, and its effectiveness is demonstrated through experiments.

In Chapter 5, latent representation-based copyright protection methods for artificially created works are introduced. An adversarial perturbation method is proposed to prevent unauthorized model training while preserving visual fidelity.

In Chapter 6, the conclusions of this thesis and directions for future research are discussed. The contributions of this thesis are summarized, and potential extensions of latent representation learning for future creative AI systems are presented.

The methods proposed in each chapter of the thesis correspond to the authors’ research achievements summarized at the end. The sections in Chapter 3 introduces the methods proposed in [A-1], [A-2], [B-1], and [B-2], respectively. These methods build upon the methods described in [C-1], [C-2] and [C-3]. Chapter 4 introduces the method proposed in [B-3], which builds upon the method described in [C-5]. Chapter 5 introduces the method proposed in [C-4].

1.4 Licensing Information for Artificially Created Works in This Thesis

Some artificially created works in this thesis are reproduced or modified from publicly available resources under Creative Commons licenses. Proper attribution is provided for each figure in accordance with the respective license terms.

Chapter 2

Related Works

2.1 Introduction

This chapter shows the research related to this thesis. Subsection 2.2 describes the related works on semantic understanding for artificially created works, subsection 2.3 shows the previous works on personalized text-to-image generation, and subsection 2.4 shows the previous works on copyright protection for artificially created works. Next, subsection 2.5 clarifies the problems to be solved in this thesis. Finally, subsection 2.6 concludes this chapter.

2.2 Semantic Understanding Methods for Artificially Created Works

Reference [16]

This paper introduces a deep residual learning framework, commonly represented by ResNet-101, which has become a fundamental backbone in many visual recognition tasks. The key idea is to employ residual connections to alleviate the gradient vanishing problem and enable stable training of very deep networks. By stacking residual blocks, the model is able to learn hierarchical visual features that capture both low-level patterns and high-level semantics. For artificially created works, such as stylized illustrations or domain-specific visual content, the hierarchical feature extraction capability of residual networks provides a strong baseline for multi-label attribute recognition. Although ResNet-101 is not explicitly designed for modeling attribute dependencies or semantic relationships, its robust feature representation has been widely adopted as a foundation for more advanced semantic understanding methods.

Reference [25]

This paper presents a method named Semantic Self-guided Representation Learning (SS-GRL), which aims to improve multi-label image classification by leveraging semantic dependencies between labels. The method introduces a self-guided attention mechanism that learns label-specific features and enhances the representation of attribute-related regions. By modeling the co-occurrence patterns and inherent correlations among attributes, SS-GRL enables more accurate recognition in domains where labels are interdependent. For artificially created works that often contain complex hierarchical attribute structures, this approach provides a useful mechanism for learning more discriminative semantic representations. The method demonstrates the importance of integrating label semantics into the feature learning process.

Reference [21]

This paper proposes a method named Deep Attentive Neural Network (DAN), that focuses on learning attention-aware representations for multi-label classification. The model incorporates an attentional mechanism that allows features to be dynamically weighted according to the relevance of each visual region to different labels. This enables the model to emphasize label-specific areas while suppressing irrelevant regions, thereby improving classification accuracy. For artificially created works with stylized or exaggerated features, attention mechanisms are effective in capturing fine-grained attribute cues that may not be uniformly distributed across the image. The approach highlights the potential of attention-based architectures for modeling semantic structures in complex visual domains.

Reference [26]

This paper introduces a framework: Multi-label Graph Convolutional Network (ML-GCN). The key contribution is the formulation of label relationships using a graph structure, where nodes represent labels and edges encode their statistical correlations. Through graph convolution, the model propagates information across related labels, allowing the classifier to explicitly leverage semantic dependencies during prediction. This graph-based approach is particularly effective for artificially created works, where hierarchical and relational attributes are essential for accurate understanding. ML-GCN demonstrates how structured

label dependencies can be incorporated into deep learning models to enhance semantic reasoning.

Reference [27]

This paper presents an Attention-Driven Dynamic Graph Convolutional Network (ADD-GCN), which extends GCN-based multi-label recognition by introducing attribute-level decomposition and cross-label dynamic feature. The method decomposes visual features into attribute-aware components and propagates them across a graph that captures semantic relations. By enabling both local and global interaction among attributes, the model achieves more consistent and interpretable predictions. For artificially created works characterized by intricate relational structures, ADD-GCN provides a more expressive representation of hierarchical dependencies compared with conventional GCN-based approaches.

Reference [28]

This paper proposes a Progressive Graph Convolutional Network (P-GCN), which models attribute relations through progressively refined graph reasoning. Instead of relying on a static correlation graph, the method iteratively updates label representations by incorporating progressively refined semantic dependencies. This dynamic reasoning process enables the model to capture both coarse and fine-grained relationships among attributes. For artificially created works that contain multilayered semantic structures, P-GCN provides an effective mechanism for learning relational features that evolve across different abstraction levels. The approach illustrates the usefulness of progressive reasoning in multi-label semantic understanding tasks.

2.3 Personalized Text-to-image Diffusion Models

Reference [10]

This paper introduces a personalization method known as Textual Inversion, which aims to represent a novel visual concept by optimizing a new token in the text embedding space of a pretrained text-to-image diffusion model. Instead of modifying model parameters, the method encodes concept-specific information into a learned embedding vector associated with a pseudo-word. This lightweight approach enables users to generate images contain-

ing personalized subjects with minimal training samples. Textual Inversion demonstrates that diffusion models possess sufficient expressive power to integrate new concepts purely through embedding optimization. However, since the learned embeddings remain entangled with the model’s existing semantic space, personalized concepts may interfere with related concepts, highlighting the challenge of maintaining clear semantic boundaries in continual learning scenarios.

Reference [7]

This paper presents a method called DreamBooth, which fine-tunes diffusion model parameters to incorporate new personalized concepts while preserving the subject’s identity. The method introduces a class-specific prior preservation loss that encourages the model to maintain diversity and avoid overfitting to a small set of training images. DreamBooth achieves high-fidelity personalization and can generate the target concept across various prompts, styles, and environments. However, the parameter-fine-tuning strategy may lead to concept bleaching or leakage of personalized information into other semantic regions of the latent space. As a result, DreamBooth can cause severe concept interference when multiple personalized concepts are learned sequentially, demonstrating limitations in continual personalization settings.

Reference [29]

This paper proposes InstantBooth, a method designed to accelerate personalization of diffusion models by integrating identity-specific embeddings with efficient fine-tuning techniques. The approach reduces computational cost while maintaining high-quality generation, making it suitable for real-world applications where rapid personalization is required. InstantBooth uses a combination of embedding optimization and partial network adaptation, enabling a better balance between specificity and generalization. Despite its efficiency, the method inherits similar challenges to other fine-tuning approaches, including the risk of concept drift and interference during continual learning, where previously learned identities may degrade when new ones are added.

Reference [30]

This paper introduces Custom Diffusion, a method that uses rank-one or low-rank param-

eter updates to enable efficient personalization. By applying concept-specific updates to attention layers, the approach allows multiple concepts to coexist without fully fine-tuning the whole model. Custom Diffusion supports the combination of concepts and reduces the risk of overwriting previous information. However, even with localized parameter updates, the latent semantic space may still experience entanglement between concepts, especially when training multiple custom concepts sequentially. This highlights the need for methods that explicitly regulate or separate latent representations during continual personalization.

Reference [31]

This paper presents an embedding-decomposed Low-Rank Adaptation (ED-LoRA) for diffusion models. ED-LoRA performs low-rank parameter adaptation while introducing mechanisms intended to reduce interference between the base model and personalized parameters. The method achieves faster training and lower memory consumption while maintaining high-quality generation. Although ED-LoRA improves parameter efficiency, the latent space of the diffusion model remains vulnerable to concept overlap in multi-concept scenarios, revealing the difficulty of maintaining robust concept separation solely through parameter-efficient fine-tuning.

Reference [23]

This paper proposes a continual LoRA-based personalization framework (C-LoRA) that enables users to combine multiple personalized concepts through continual controls. By decoupling personalized concepts into separate LoRA modules, the method allows flexible concept mixing and reduces the risk of overwriting. C-LoRA demonstrates that modularization is an effective way to enhance conceptual flexibility in diffusion models. However, when concepts are learned sequentially, interference between LoRA modules can still occur through shared latent representations, suggesting that modularity alone is insufficient for addressing concept bleed-through in continual learning settings.

Reference [2]

This paper introduces a personalization method that employs Stack-and-mask Incremental Adapters (STAMINA) to maintain concept integrity during multi-concept learning. The method selectively updates model parameters that are most relevant to each concept, re-

ducing overlap in parameter usage across concepts. This improves concept retention and alleviates interference during sequential learning. However, the latent semantic representations of concepts may still interact through shared diffusion pathways, indicating that parameter sparsity must be complemented by latent-space regulation to achieve robust concept disentanglement.

Reference [24]

This paper presents a Concept-Incremental text-to-image Diffusion Model (CIDM) designed to mitigate forgetting when incorporating new personalized concepts. The method adopts identity-preserving constraints and replay mechanisms to stabilize previously learned concepts while adding new ones. CIDM improves continual learning performance by reducing degradation of older identities. Nevertheless, as the number of concepts increases, the replay mechanism becomes less effective, and entanglement in the latent space may still result in concept bleed-through. This highlights the fundamental difficulty of maintaining clear semantic boundaries within the diffusion latent space under continual personalization.

2.4 Copyright Protection Methods for Artificially Created Works

Reference [3]

This paper introduces Glaze, a protection method designed to prevent unauthorized style extraction by generative models. Glaze applies subtle perturbations to input artworks so that the perturbed images appear unchanged to human observers but differ significantly in the latent space of deep generative models. As a result, models trained on protected images fail to capture the original artistic style, thereby reducing the risk of stylistic imitation. The method demonstrates that perceptually aligned but semantically disruptive perturbations can serve as an effective defensive strategy against unpermitted dataset scraping. However, because the perturbations primarily target the feature space of specific model families, the method may exhibit reduced robustness against unseen architectures or future-generation diffusion models, indicating a need for more generalized protection mechanisms.

Reference [4]

This paper presents an adversarial perturbation-based method designed to mislead diffu-

sion models (AdvDM). AdvDM optimizes image perturbations that specifically disrupt the denoising process used in diffusion-based generative models, preventing them from learning meaningful style or content representations. By directly targeting the generative pathway, the method significantly reduces the ability of diffusion models to reproduce protected artworks. Although AdvDM achieves strong protective effects, it may produce artifacts under certain perturbation strengths, and its effectiveness depends on the generalization behavior of the targeted diffusion architecture. These limitations highlight the difficulty of designing protection methods that maintain both high visual fidelity and long-term robustness.

Reference [5]

This paper introduces Anti-DreamBooth, a protection strategy aimed at preventing personalized concept learning, particularly against methods such as DreamBooth that fine-tune a diffusion model on a small set of images. Anti-DreamBooth generates perturbations that cause fine-tuned models to fail to reproduce the protected concept, even when the attacker uses multiple training samples. By disrupting the latent-space consistency required for personalization, the method effectively prevents the model from capturing identity-related or style-related information. Although Anti-DreamBooth successfully blocks many personalization attacks, its performance varies depending on the attacker’s training configuration and may be less effective when the adversary employs enhanced optimization strategies. This suggests the need for more adaptive protection mechanisms that remain robust across diverse personalization scenarios.

2.5 Problems to Be Solved in This Thesis

Although significant progress has been made in semantic understanding, personalized text-to-image diffusion modeling, and copyright protection, existing methods still face fundamental limitations when applied to artificially created works. First, methods for semantic understanding rely heavily on hierarchical or graph-based reasoning, yet they often struggle to capture the fine-grained and domain-specific attribute structures present in stylized or nonphotorealistic illustrations. These methods lack explicit mechanisms for modeling the latent semantic relationships

characteristic of artificially created works, resulting in limited interpretability and insufficient representation capacity.

Second, personalization methods for diffusion models, including embedding-based, fine-tuning-based, and modular approaches, exhibit difficulty in maintaining stable concept boundaries during continual learning. Concept bleed-through remains a central challenge, as newly introduced concepts frequently interfere with previously learned ones due to entanglement in the latent space. Existing approaches primarily modify model parameters or embeddings but do not directly address latent-space regulation, which is essential for achieving robust and disentangled personalization.

Third, copyright protection methods based on adversarial perturbations successfully disrupt generative model training but often lack generalization across different architectures or training configurations. Their effectiveness tends to degrade when adversaries employ model adaptation, fine-tuning, or defensive preprocessing techniques. Additionally, these methods do not explicitly consider latent semantic structures, resulting in protection mechanisms that are fragile or overly dependent on specific model families.

To address these limitations, this thesis focuses on latent representation learning as a unified framework for understanding, generating, and protecting artificially created works. By explicitly modeling and regulating latent semantic structures, the goal of this thesis is to achieve improved semantic understanding, robust generative controllability, and effective copyright protection within a coherent and integrated approach. Figure 2.1 shows a research map of related works that summarizes the above, and the position of each chapter in this thesis.

2.6 Conclusion

This chapter explained the research related to this thesis. Furthermore, the problems to be solved in this thesis are clarified.

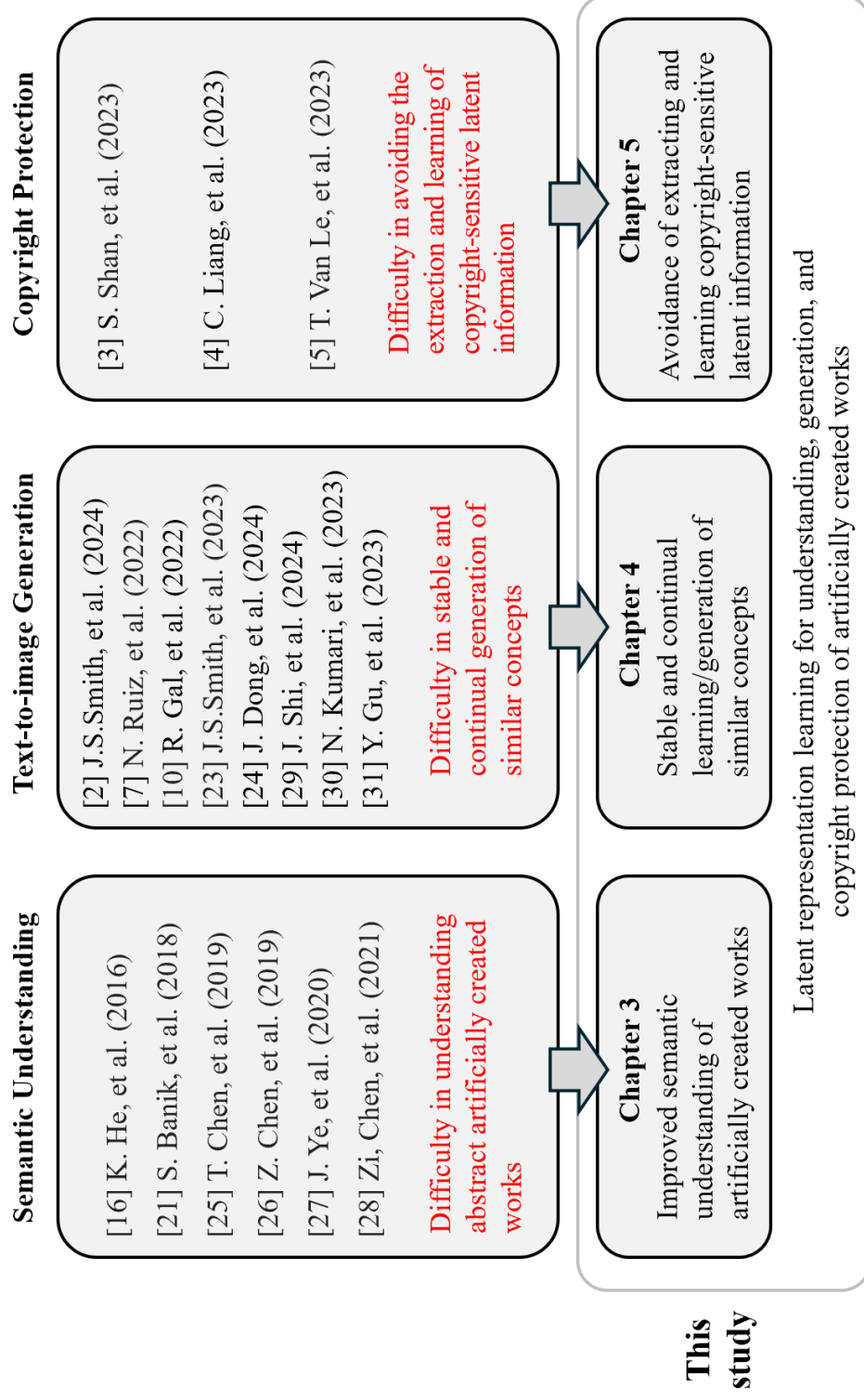


Figure 2.1: Research map of related works and position of each chapter in this thesis.

Chapter 3

Latent Representation-based Semantic Understanding Methods for Artificially Created Works

Recent advances in deep generative models and large-scale visual–language learning have significantly expanded the scope of semantic understanding for images. Traditional image understanding methods primarily rely on predefined labels or handcrafted annotations, which often fail to capture the rich, compositional semantics present in complex visual content. To address these limitations, image captioning has emerged as a promising paradigm by providing natural language descriptions that bridge visual representations and semantic concepts. Leveraging latent representations derived from image captioning models offers a unified and flexible approach to semantic understanding across diverse types of visual data.

The first line of research focuses on real-world images, where visual content is inherently complex, noisy, and semantically diverse. In this context, latent representation–based semantic understanding using image captioning enables models to encode high-level semantic information beyond object-level recognition. By mapping images into a latent semantic space guided by generated captions, this approach allows for robust understanding of scenes, activities, and contextual relationships. Such methods demonstrate that latent representations learned through captioning can effectively capture semantic similarity and abstraction in real-world imagery, providing a foundation for downstream tasks such as retrieval, classification, and semantic comparison.

Building upon this framework, the second study extends the latent representation–based semantic understanding method to artificially created works, such as illustrations, digital art, or synthetic images. Unlike real-world images, artificially created works often exhibit exaggerated styles, symbolic elements, or non-photorealistic structures, posing challenges for conventional vision models trained on natural image distributions. By applying image captioning–based latent representations, this research shows that semantic understanding can be generalized beyond real-world imagery. The generated captions serve as an intermediate semantic bridge, enabling the model to interpret abstract or stylized visual elements in a manner consistent with human perception. This extension highlights the adaptability of latent semantic representations and underscores their potential for understanding a broader spectrum of visual content.

However, while latent representations provide powerful semantic abstraction, artificially created works often involve hierarchically structured concepts, such as genre, theme, style, and sub-style, which are not explicitly captured by flat label representations. To address this limitation, the third study introduces label hierarchy into the semantic understanding framework. By incorporating hierarchical label structures into the learning process, the proposed method aligns latent representations with multi-level semantic relationships. This hierarchical consideration enables finer-grained semantic discrimination while preserving high-level conceptual coherence. As a result, the model can better reflect structured semantic knowledge and improve interpretability and robustness in understanding artificially created works.

Taken together, these three studies form a coherent progression from semantic understanding of real-world images to artificially created works, and finally to hierarchical semantic modeling. They collectively demonstrate that image captioning–based latent representations provide a versatile and extensible foundation for semantic understanding across diverse visual domains. By progressively addressing domain generalization and semantic structure, this line of research contributes toward more flexible, interpretable, and human-aligned visual understanding systems.

The remainder of this chapter is organized as follows. Section 3.1 introduces a latent representation–based semantic understanding method for real-world images using image captioning, focusing on how caption-guided latent spaces enable robust semantic abstraction in complex natural scenes. Section 3.2 extends this framework to artificially created works, addressing the challenges posed by stylized, symbolic, and non-photorealistic visual content and demonstrating

the generalizability of the proposed semantic representation. Section 3.3 further incorporates label hierarchy into the latent semantic learning process, enabling structured and multi-level semantic understanding for artificially created works and improving both discriminability and interpretability.

Chapter 3.1

Latent Representation-based Semantic Understanding Method for Real-World Images Using Image Captioning

3.1.1 Introduction

In the computer vision community, image recognition is a fundamental task and plays an important role in wide applications. Depending on the content of an image, we use labels to describe it. Since nature scenes often contain multiple categories of objects, multi-label recognition has attracted much attention. It can tell us whether an image contains multiple contents at the same time, which can also be a better solution to real-life problems such as human attribute recognition [32].

Since objects are usually co-occurring in the physical world, it is necessary to consider the relationships between different labels to improve the recognition performance. In recent years, along with the progress of research on Graph Convolutional Networks (GCN) [33], it has achieved great success in modeling the relationship between labels. For example, Chen et al. proposed a GCN-based multi-label recognition model (ML-GCN) which uses the label co-occurrence prior frequency of the dataset [26]. They modeled the label correlation between each of the two categories by building a complete graph. It shows significant advantages in scalability and flexibility that previous approaches did not have. However, if we build a global graph for the whole dataset like this, frequency bias problems may occur in commonly used datasets. To solve this problem, Ye et al. proposed an Attention-Driven Dynamic Graph Convolutional Network (ADD-

GCN) [27]. In this model, a specific graph is dynamically generated for each image to capture the label relationships of perceptible contents of each image.

The aforementioned approaches focused on the capture of relational properties between labels from the image. However, it is more accurate to label images with both images and texts. Images provide comprehensive information while text summarizes descriptive features of objects, and both types of information help people recognize objects. Therefore, the image captioning [34], which is a more exhaustive factor consisted of the descriptions, can be expected to be effective for image recognition. By introducing image captioning into the GCN model, there is a great possibility to improve the performance of capturing the relationships of the labels.

Therefore, we propose a multi-label image recognition model based on multi-modal GCN using captioning features. We extract the captioning features by using pre-trained captioning model and merge the captioning features with the visual features generated by the ADD-GCN model. The merged features are imported into the dynamic graph convolution network. As a result, the improvement of multi-label image recognition is achieved. The experiments show that our multi-label image recognition model achieves a better performance after the introduction of the captioning feature.

3.1.2 Proposed Method

This section explains the details of our model. We describe the structure of ADD-GCN and then explain how to import captioning features into it. The overview of our model is shown in Fig.3.1.1.

3.1.2.1 ADD-GCN

ADD-GCN [27] is mainly composed of a Semantic Attention Module (SAM) and a Dynamic GCN Module (DGCNM). We estimate the content-aware category representation from the feature maps extracted from the images by SAM, and feed that representation into the DGCNM to obtain the final category representation using the content of each particular input image for final classification.

The SAM aims to get a set of content-aware category representations, and each of them describes the content associated with a particular label from the input feature map $X \in \mathbb{R}^{H \times W \times D}$.

H , W and D are the height, width and number of channels of X . We generate category-specific activation maps $\mathbf{M} \in \mathbb{R}^{H \times W \times C}$ from the input feature map X by using the Class Activation Mapping (CAM) [35], and convert X into a category representation of the perceptible content $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_C]^\top \in \mathbb{R}^{C \times D}$ as follows:

$$\mathbf{v}_c = \sum_{i=1}^H \sum_{j=1}^W m_{i,j}^c \mathbf{x}_{i,j}, \quad (3.1.1)$$

where $m_{i,j}^c$ and $\mathbf{x}_{i,j} \in \mathbb{R}^D$ are the weight of c -th activation map $\mathbf{M}_c \in \mathbb{R}^{H \times W}$ and the feature vector of the feature map X at $(i; j)$, respectively. C is the number of categories.

We obtain the content-aware category representation $\mathbf{V}$ and then introduce it into the DGCNM. The category representation $\mathbf{V}$ is used as features of the input node and is given sequentially to a static GCN and a dynamic GCN. We simply define the single-layer static GCN as $\mathbf{H} = LReLU(\mathbf{A}_s \mathbf{V} \mathbf{W}_s)$, where $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_C]^\top \in \mathbb{R}^{C \times D}$, and the activation function $LReLU(\cdot)$ is LeakyReLU [36]. $\mathbf{A}_s$ is the correlation matrix, and $\mathbf{W}_s$ is the state update weights. Then the dynamic GCN is adopted. The correlation matrix $\mathbf{A}_d$ is adaptively estimated based on a feature matrix $\mathbf{H}'$ which is affected by the captioning features. The details are explained in II-B. The output $\mathbf{Z} \in \mathbb{R}^{C \times D}$ of the DGCNM can be defined as follows:

$$\mathbf{Z} = LReLU(\mathbf{A}_d \mathbf{H} \mathbf{W}_d), \text{ where } \mathbf{A}_d = \sigma(\mathbf{W}_A \mathbf{H}'). \quad (3.1.2)$$

$\mathbf{W}_d$ is the state-update weights, $\sigma(\cdot)$ is the Sigmoid activation function, and $\mathbf{W}_A$ is the weights of a *conv* layer to formulate the dynamic correlation matrix $\mathbf{A}_d$.

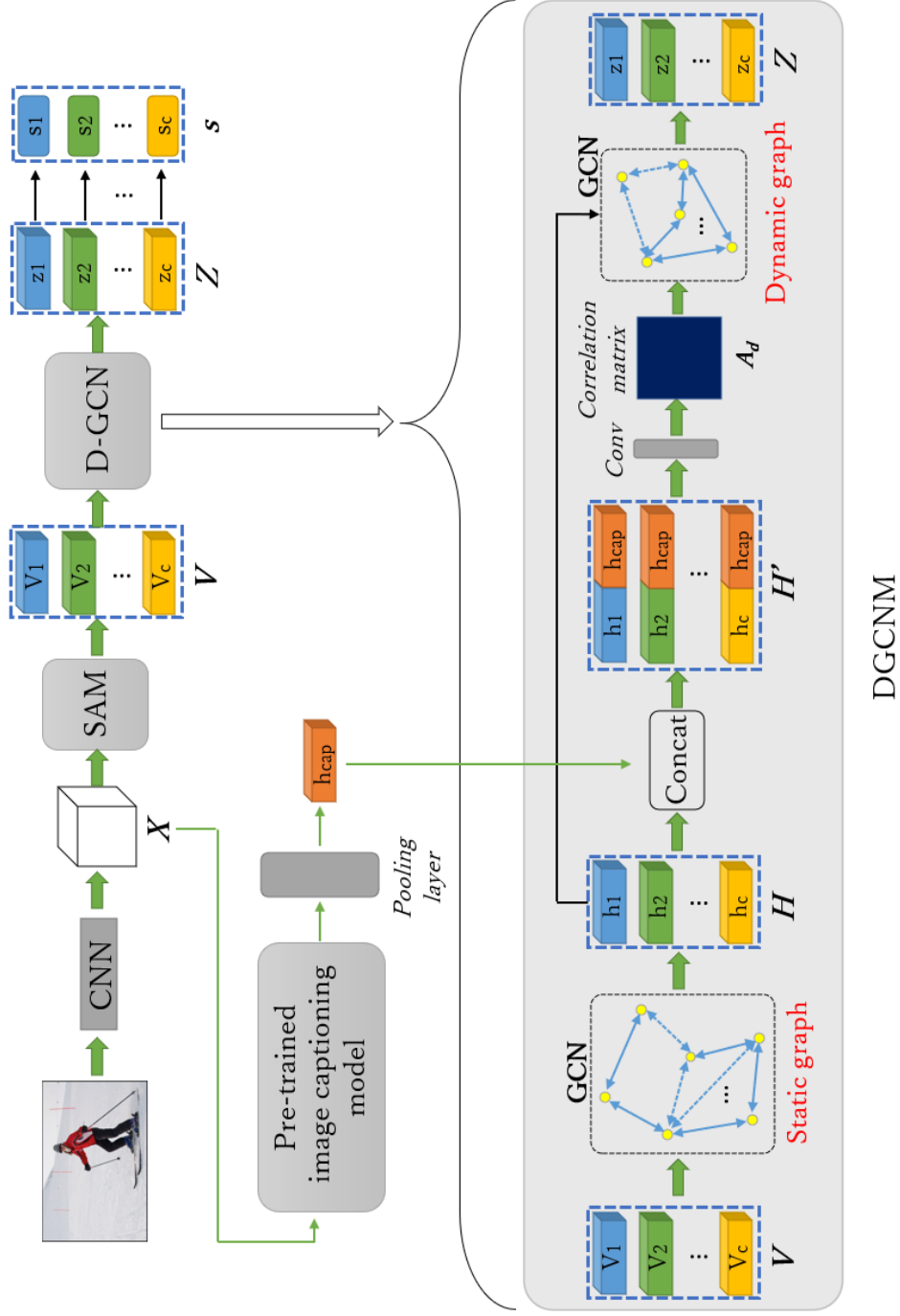


Figure 3.1.1: Overview of our model. The captioning features from the pre-trained captioning model are processed by the pooling layer and then combined with the output H of the static GCN network in the ADD-GCN model.

3.1.2.2 Import of Captioning Features

To get the captioning features, we extract the captioning matrices $X_{cap} \in \mathbb{R}^{H \times W \times D}$ representing the image content information by employing a pre-trained image captioning model. In order to import the captioning feature into the ADD-GCN model, we use a pooling layer to transform the captioning matrices into captioning features $\mathbf{h}_{cap} \in \mathbb{R}^D$ that can be combined with the output $\mathbf{H}$. Captioning features $\mathbf{h}_{cap}$ is defined as follows:

$$\mathbf{h}_{cap} = p(X_{cap}), \quad \text{s.t. } X_{cap} = M_{cap}(X), \quad (3.1.3)$$

where p is the global average pooling layer, and M_{cap} is the pre-trained captioning model. Then we concatenate $\mathbf{h}_c$ ($c = 1, 2, \dots, C$) with $\mathbf{h}_{cap}$ to obtain the combined feature matrix $\mathbf{H}' = [(\mathbf{h}_1; \mathbf{h}_{cap}), (\mathbf{h}_2; \mathbf{h}_{cap}), \dots, (\mathbf{h}_C; \mathbf{h}_{cap})]^\top \in \mathbb{R}^{C \times 2D}$. This matrix $\mathbf{H}'$ contains rich label correlations and captioning features derived from the image, and it can replace $\mathbf{H}$ in the original ADD-GCN model for the same dynamic GCN transformation.

3.1.2.3 Final Classification and Loss

The output $\mathbf{Z} = [z_1, z_2, \dots, z_C]^\top$ of the DGCNM is considered as the final category representation, and is used for final classification. To predict the scores of each category, each category vector is simply put into a binary classifier. In particular, the final score vector $\mathbf{s} = [s^1, s^2, \dots, s^C]^\top$ is generated by concatenating the scores of each category. We supervise the final score $\mathbf{s}$ and train the model with the traditional multi-label classification loss [26] as follows:

$$L(\mathbf{y}, \mathbf{s}) = \sum_{c=1}^C y^c \log(\sigma(s^c)) + (1 - y^c) \log(1 - \sigma(s^c)), \quad (3.1.4)$$

where $\sigma(\cdot)$ is the Sigmoid function, $\mathbf{y} \in \mathbb{R}^C$ is the ground truth label of an image, and $y^c = \{0, 1\}$ denotes whether label c appears in the image or not.

3.1.3 Experiments

In our experiments, we used the MS COCO [37] dataset for training. The MS COCO dataset consists of a training set with 82,081 images and a validation set with 40,137 images. It includes 80 common object classes and each image contains an average of 2.9 object labels. We used

Table 3.1.1: Comparison of our model and other models. The best results are marked as bold.

Method	CP	CR	CF1	OP	OR	OF1	mAP
ML-GCN [26]	85.1	72.0	78.0	85.8	75.4	80.3	83.0
ADD-GCN [27]	84.7	73.6	78.9	86.2	76.6	81.1	84.2
Ours	91.3	71.2	80.2	89.8	74.1	85.4	85.9

ResNet-101 as the backbone of the ADD-GCN framework. We randomly cropped the input images, resize them to 448×448 and flipped them horizontally for data augmentation to avoid over-fitting. In addition, we employed the image captioning model [34] pre-trained with the MS COCO dataset. We compared our method with other state-of-the-art methods. In particular, We adopted ML-GCN [26] and ADD-GCN [27] as comparison methods.

To compare with other method, we followed previous studies [26] [27] and adopted the average of per-class precision (CP), per-class recall (CR), per-class F1 score (CF1), overall precision (OP), overall recall (OR), overall F1 score (OF1), and the mean Average Precision (mAP) as evaluation metrics. After sorting the confidence score in descending order, the mAP is calculated. When measuring precision/recall/F1-score, we classify a label as positive if its confident score is higher than 0.5. As shown in Table 3.1.1, it can be seen that the performance of the model after the introduction of captioning features is better than other models in the comparison of several metrics mentioned above. It can be proved that our method outperforms other methods. Figure 3.1.2 shows an example of ADD-GCN and our method. As shown in this figure, the effectiveness of introduction of captioning features is verified.

3.1.4 Conclusions

The multi-label image recognition model based on ADD-GCN using captioning features has been proposed. It is shown that our model has a better performance.

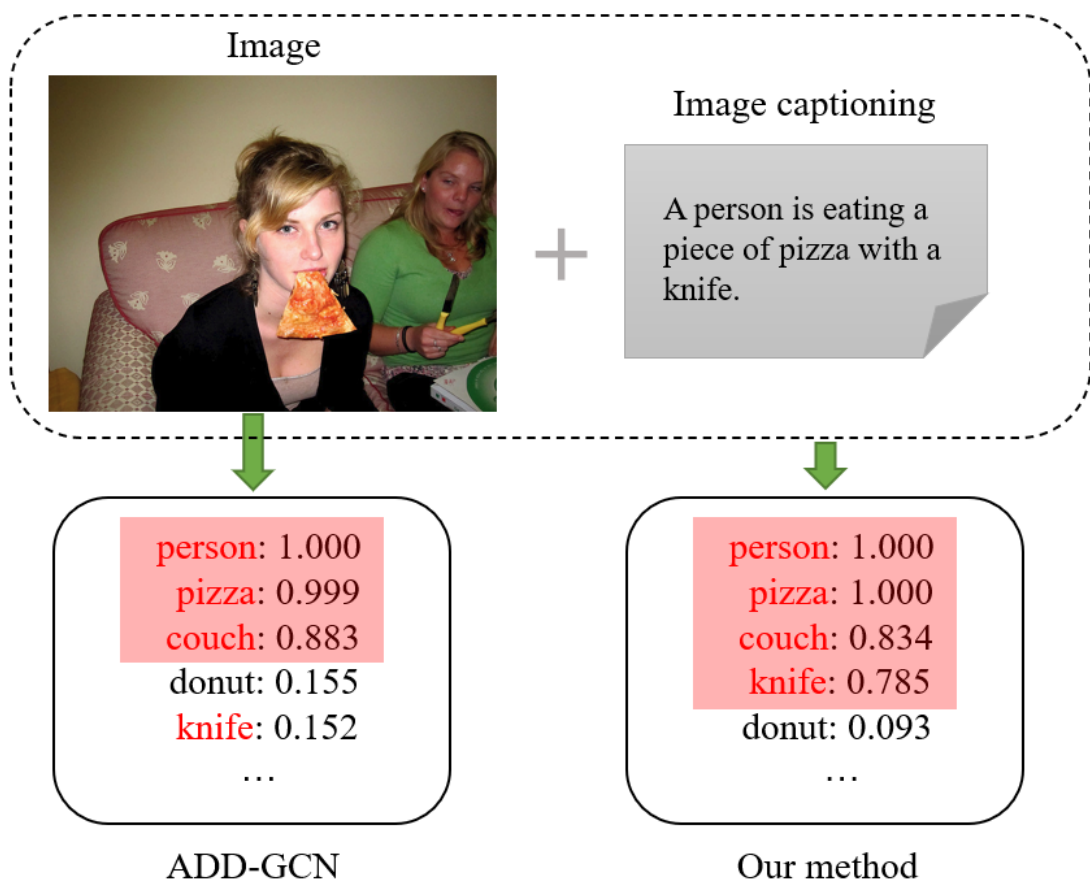


Figure 3.1.2: An example of ADD-GCN and our method. All the confidence scores are sorted in descending order. Labels in red font are the ground truth labels, and labels in the red box represent the final classification.

Chapter 3.2

Latent Representation-based Semantic Understanding Method for Artificially Created Works Using Image Captioning

3.2.1 Introduction

With the development of the animation industry in recent years, several studies related to anime illustration such as illustration editing [38, 39], line-art colorization [40, 41] and cartoon face generation [42, 43] have been conducted. Due to the increasing number of anime illustrations, there is a growing need for techniques to classify them, and these techniques have a potential to enhance the above various studies. In order to implement the classification techniques for anime illustration, it is necessary to know the contents of the anime illustrations.

It is well known that image classification is a fundamental task in the computer vision community and plays an important role in a wide range of applications. Since images contain multiple objects in the great majority of cases, multi-label image classification has aroused extensive attention in recent years. In real-life problems such as medical image classification [44] and recommendation systems [45], multi-label image classification can help us get better solutions. Therefore, it can be expected that the multi-label classification technique is also effective for anime illustrations.

Many multi-label classification methods for anime illustration have been proposed [1, 46]. For

example, a method [1] based on convolutional neural network (CNN) achieved good classification results. Since objects always co-occur in anime illustrations, it is necessary to take the relationships between labels into consideration so that we can improve the accuracy of classification. In recent years, great progress has been made in the research of graph convolutional network (GCN) [33]. GCN shows a great ability for modeling the relationships between the labels on a graph structure. For example, Feng et al. proposed a GCN-based multi-label anime illustration classification method [46], which modeled the correlation between different labels by constructing a complete graph.

However, the classification task of anime illustration is slightly different from that of real-world image. Anime illustrations are artificially created images, and the creators of anime illustrations often intentionally highlight the fine characteristics of the characters and objects which are called attributes. For this reason, to construct the classification method for anime illustrations, we need to consider not only simple objects in the image but also their attributes. Therefore, it is necessary to solve the multi-label attribute classification task [21], and there is no work focusing on this task for anime illustration.

In generic image classification, many methods for multi-label attribute classification have been proposed [47–49], and they focused on the capture of relationship between attributes from the visual information of the illustration. As mentioned above, artificially created anime illustrations contain more detailed information than real-world images, so it is necessary to take more semantic information into account. However, it is difficult to grasp the deeper and more detailed connections between attributes only from visual information. To grasp semantic features specific to anime illustration, we introduce the image captioning [50] into the classification task of anime illustration. The image captioning is a more detailed factor consisted of the semantic information, and is expected to be useful for representing the attributes. However, the available models [34, 50] of image captioning are learned from only real-world images. There are several differences between anime illustrations and real-world images which are called domain-shift, and the domain-shift will lead to a decrease in the accuracy of image captioning results. In other words, it is not feasible to apply this technique for real-world images directly to anime illustrations.

Therefore, we propose a multi-modal multi-label attribute classification model for anime il-

illustrations based on GCN using domain-specific semantic features. This is the first attempt to introduce the semantic features to the task of multi-label attribute classification in anime illustrations. The technical contributions of this method are as follows.

Contribution (i): To deal with the task, multi-label attribute classification in anime illustration, we construct a GCN-based model using not only visual features from images but also semantic features suitable for anime illustration.

Contribution (ii): We propose a trainable model for the image captioning of anime illustrations by using only real-world images to extract highly expressive image captioning of anime illustrations.

In contribution (i), we can capture the relationship between different attributes in the images by introducing semantic features for anime illustration based on image captioning into GCN, and we further improve the classification accuracy. In contribution (ii), we convert real-world images into animated style images and import them in pairs into the image captioning model pre-trained with real-world images, and train the model to make each feature space closer by considering the differences between the two captioning results. In this way, we can generate captioning of anime illustrations by using only real-world images. These two contributions show that successful multi-label attribute classification for anime illustration becomes feasible.

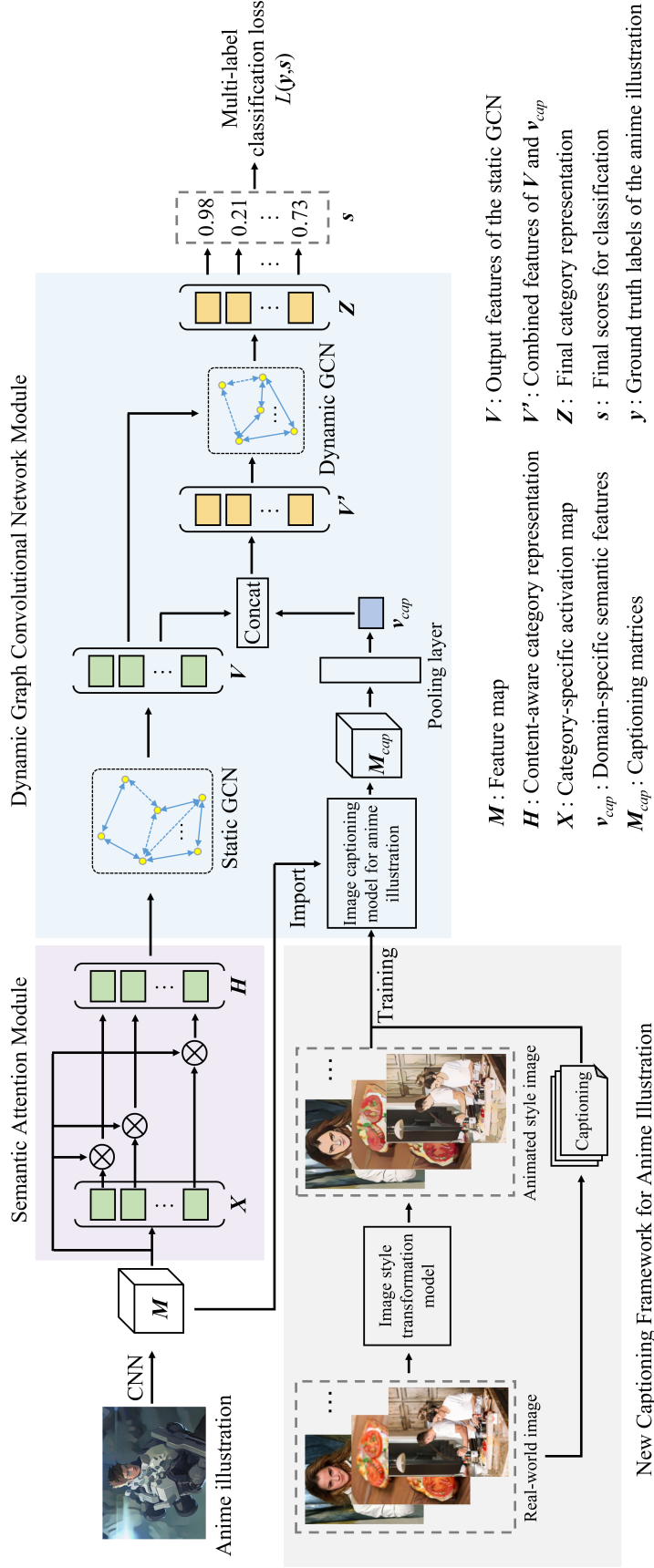


Figure 3.2.1: Overview of the GCN-based multi-modal attribute classification model for anime illustrations using domain-specific semantic features. We train a captioning model for anime illustration by using the captioning of real-world images and corresponding animated style images generated by the image style transformation model. We employ this captioning model to obtain the semantic feature v_{cap} of anime illustration. The semantic feature v_{cap} is used in the GCN-based multi-label classification model to improve the classification accuracy. Source: Anonymous authors, <https://pixabay.com>. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). Modified by the author.

3.2.2 Proposed Method

In this section, we show the details of the proposed model. First, we explain how we construct a new captioning framework for anime illustrations (3.2.2.1). Then we describe the general flow of the GCN-based multi-modal multi-label attribute classification using domain-specific semantic features (3.2.2.2). Finally, we explain the final classification and the loss used for the training of the proposed model (3.2.2.3). Figure 3.2.1 shows the overview of the proposed model.

3.2.2.1 New Captioning Framework for Anime Illustration

This subsection shows how we construct the new captioning framework for anime illustrations by combining real images and their corresponding art style transformations to generate features that represent the semantic contents of anime illustrations.

First, we put real-world images into the image style transformation model and convert them into animated style images. In our method, we use white-box cartoonization model [51] as the image style transformation model. The white-box cartoonization model can generate high-quality cartoonized images from real-world images, and with this model, we can get pairs of real-world images and animated style images.

Next, we input real-world images into a pre-trained captioning model [34] to obtain the corresponding captioning results. Under normal circumstances, since simply converting a real image into an animated style image does not change rough contents of the image, we can regard the captioning of the real-world image as the ground truth of that of the corresponding animated style image. We use the animated style image and the captioning of the corresponding real-world image to train the captioning model so that the distance between the captioning result of the animated style image and that of the real-world image becomes close. The training process will be mentioned in Section 3. In this way, the problem of the domain-shift can be solved, and we can obtain domain-specific semantic features for anime illustrations.

3.2.2.2 GCN-based Multi-modal Multi-label Attribute Classification Using Domain-specific Semantic Features

This subsection describes the general flow of the GCN-based multi-label attribute classification using domain-specific semantic features. Inspired by the uni-modal version of previous GCN [27], our multi-label attribute classification model mainly consists of the following two parts: Semantic Attention Module (SAM) and Dynamic Graph Convolutional Network Module (DGCNM) as shown in Fig. 1.

The SAM aims to estimate the content-aware category representations after extracting feature maps from the images. For an input image, we first employ a CNN backbone to extract the convolutional feature maps $\mathbf{M} \in \mathbb{R}^{H \times W \times D}$. H means the height of $\mathbf{M}$, W means the width of $\mathbf{M}$, and D means the number of channels of $\mathbf{M}$. From the feature maps $\mathbf{M}$, a set of category-specific activation maps $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ are generated by Class Activation Mapping [35]. Then we use the activation maps $\mathbf{X}$ to convert $\mathbf{M}$ into a content-aware category representation $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_C]^\top \in \mathbb{R}^{C \times D}$ as follows:

$$\mathbf{h}_c = \sum_{i=1}^H \sum_{j=1}^W x_{i,j}^c \mathbf{m}_{i,j}, \quad (3.2.1)$$

where C is the number of categories, $x_{i,j}^c$ is the (i, j) -th weight of c -th activation map $\mathbf{X}_c \in \mathbb{R}^{H \times W}$, and $\mathbf{m}_{i,j} \in \mathbb{R}^D$ is the (i, j) -th feature vector of the feature map $\mathbf{M}$.

The DGCNM aims to obtain the final category representation by GCN using the contents in each specific input image. We take the content-aware category representation $\mathbf{H}$ as the input node features, and feed it into a static GCN. We define the single-layer static GCN as $\mathbf{V} = LReLU(\mathbf{A}_s \mathbf{H} \mathbf{W}_s) \in \mathbb{R}^{C \times D}$. $LReLU(\cdot)$ is the activation function LeakyReLU [36]. $\mathbf{A}_s$ and $\mathbf{W}_s$ are respectively the correlation matrix and the state update weights. During the training process, $\mathbf{A}_s$ and $\mathbf{W}_s$ are randomly initialized by the gradient descent.

To introduce the semantic features into the GCN model, we employ the image captioning model that is trained for anime illustration in the previous subsection to obtain the captioning matrices $\mathbf{M}_{cap} \in \mathbb{R}^{H \times W \times D}$. Then we transform the captioning matrices into semantic features $\mathbf{v}_{cap} \in \mathbb{R}^D$ that can be imported into the multi-label classification model by a pooling layer. Note

that $\mathbf{v}_{cap}$ is defined as follows:

$$\mathbf{v}_{cap} = p(\mathbf{M}_{cap}), \quad \text{s.t. } \mathbf{M}_{cap} = cap(\mathbf{M}), \quad (3.2.2)$$

where $p(\cdot)$ means the pooling layer, and $cap(\cdot)$ means the captioning model mentioned in subsection 2.1. After that, we integrate $\mathbf{v}_{cap}$ and the output features $\mathbf{V}$ of the static GCN. Specifically, we simply concatenate $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_C]^\top \in \mathbb{R}^{C \times D}$ with $\mathbf{V}_{cap} = [\mathbf{v}_{cap}, \mathbf{v}_{cap}, \dots, \mathbf{v}_{cap}]^\top \in \mathbb{R}^{C \times D}$ to get the combined feature $\mathbf{V}' = [\mathbf{V}, \mathbf{V}_{cap}] \in \mathbb{R}^{C \times 2D}$. Then we import $\mathbf{V}'$ into the dynamic GCN. We define the output $\mathbf{Z} \in \mathbb{R}^{C \times D}$ of the dynamic GCN as follows:

$$\mathbf{Z} = LReLU(\mathbf{A}_d \mathbf{V} \mathbf{W}_d), \quad \text{where } \mathbf{A}_d = \sigma(\mathbf{W}_A \mathbf{V}'). \quad (3.2.3)$$

$\mathbf{W}_A$ and $\mathbf{W}_d$ mean the the weights of a *conv* layer that formulates the dynamic correlation matrix $\mathbf{A}_d$ and the state-update weights, respectively. $\sigma(\cdot)$ represents the sigmoid activation function. As a result, we can capture the relationship between different attributes in the images by GCN and import the information of textual descriptions of the anime illustration into the GCN to further improve the classification accuracy.

3.2.2.3 Final Classification and Loss

This subsection explains the final classification and loss. For the final classification, we use the output $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_C]^\top$ of the dynamic GCN and predict scores of each category. Particularly, we put each vector $\mathbf{z}_c$ of the final category representation $\mathbf{Z}$ into a binary classifier and get the predict scores s^c of category c . We concatenate the scores s^c to obtain the final scores $\mathbf{s} = [s^1, s^2, \dots, s^C]^\top$. According to previous studies [26–28], the loss function $L(\mathbf{y}, \mathbf{s})$ is calculated as follows:

$$L(\mathbf{y}, \mathbf{s}) = \sum_{c=1}^C y^c \log(\sigma(s^c)) + (1 - y^c) \log(1 - \sigma(s^c)), \quad (3.2.4)$$

where $\mathbf{y} \in \mathbb{R}^C$ means the ground truth labels of an image, and $y^c = \{0, 1\}$ indicates the presence or absence of label c in the image.

3.2.3 Experiments

3.2.3.1 Training of Captioning Model for Anime Illustrations

In order to perform the transfer learning of the captioning model, we first need to produce a dataset of animated style images from real-world images. We used 8,000 real-world images from Flickr8k dataset [52] and imported them into the image style transformation model [51] shown in subsection 3.2.2.1 to generate their corresponding animated style images. Then we used the animated style images as input and the captioning of the corresponding real-world images as ground truth to train the captioning model [34].

3.2.3.2 Training of Whole Multi-label Classification Model

In our experiments, we used Danbooru2020 [1] dataset for training the whole multi-label attribute classification model. Danbooru2020 dataset is a large anime illustration dataset with over 4.2 million images and over 130 million tags. From the dataset, we extracted about 25,000 anime illustrations, which include 100 common attribute classes, and each illustration contains an average of 6.3 attribute labels. We used 75% of the 25,000 images as the training set and the remaining 25% as the validation set.

We employed ResNet-101 [16] as the backbone of the GCN-based attribute classification model. The negative slope of LeakyReLU utilized in the DGCNM was set to 0.2. The input images were randomly cropped, resized to 448×448 pixels and flipped horizontally for data augmentation. We chose Stochastic Gradient Descent as our optimizer. Its momentum and weight decay were respectively set to 0.9 and 1.0×10^{-4} . We set the initial learning rate to 0.5 for SAM/DGCNM and 0.05 for the backbone CNN.

3.2.3.3 Comparison Methods and Evaluation Metrics

To verify the effectiveness of the proposed method, we employed a series of comparison methods as follows: ResNet-101 [16], Semantic Specific Graph Representation Learning (SS-GRL) framework [25], Deep Attribute Network (DAN) [21], Multi-Label Image Recognition with GCN (ML-GCN) [26], Attention-Driven Dynamic GCN (ADD-GCN) [27] and Prediction Learning GCN (P-GCN) [28]. Comparison method ResNet-101 is the backbone of our method.

Table 3.2.1: Performance comparison between our model and other image classification models. We mark the best results as bold.

Method	OP	OR	OF1	CP	CR	CF1	mAP
ResNet-101 [16]	60.8	55.4	58.3	62.0	58.0	61.3	63.4
SSGRL [25]	64.0	56.6	61.1	71.2	57.2	60.2	69.4
DAN [21]	64.7	51.8	57.5	67.1	56.6	61.3	64.3
ML-GCN [26]	61.2	53.6	59.1	68.8	61.4	65.9	66.3
ADD-GCN [27]	63.4	59.2	60.3	71.6	63.1	68.4	70.1
P-GCN [28]	65.3	57.8	60.7	73.9	58.5	69.8	70.4
Ours	67.6	58.1	61.4	73.0	63.8	71.0	71.2

Table 3.2.2: The ablation experimental results. We can see the influence of the accuracy of the captioning model (BLEU) on the final classification (mAP). In the ablation experiment, we randomly selected 100 anime illustrations with ground truth captionings from Danbooru2020 dataset and calculated BLEU score and mAP.

Method	BLEU [14]	mAP
ADD-GCN + captioning (by model trained for real image)	44.5	64.1
Ours (using captioning obtained by model trained for anime illustration)	54.2	67.4

SSGRL and DAN are the classification models based on CNN, and ML-GCN, ADD-GCN and P-GCN are the classification models based on GCN. Among these comparison methods, ADD-GCN and P-GCN are the state-of-the-art methods. Each of these models was trained with the same hyper-parameters as ours. To confirm the performance improvement after employing the image captioning in multi-label classification model, all the comparison methods were trained with visual features of the anime illustrations only.

To compare our method with other methods, according to previous studies [26–28], we adopted the average of overall precision, recall, F1 score (expressed respectively as OP, OR and OF1) and the average of per-class precision, recall, F1 score (expressed respectively as CP, CR and CF1) to evaluate the performance of the methods. When measuring the precision, recall and F1 score, for each image, the label c is predicted as positive if the score s^c calculated in subsection 2.4 is greater than 0.5. Also, we adopted the average precision (AP) and the mean average precision

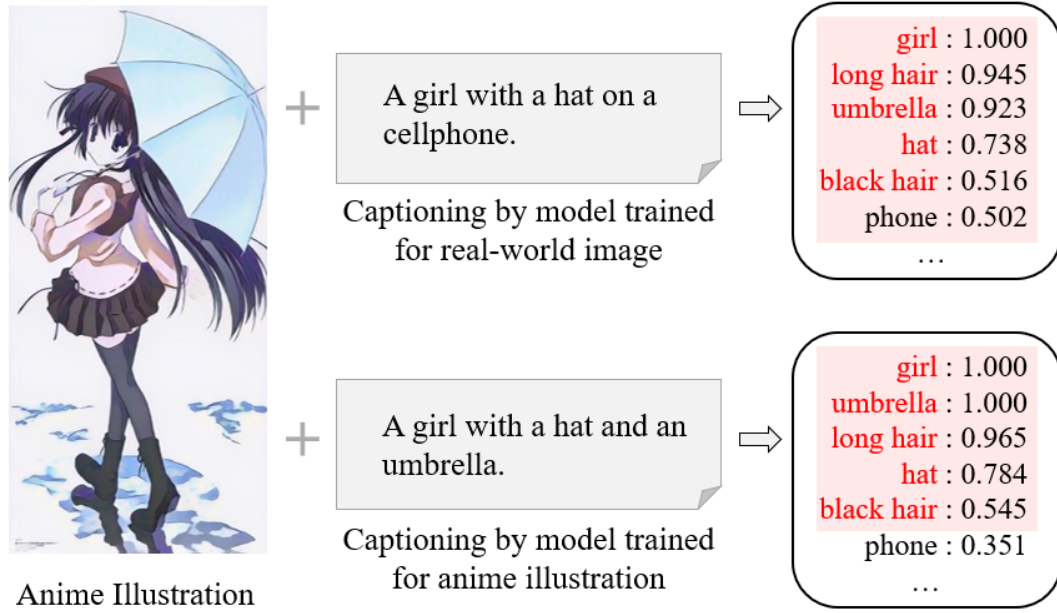


Figure 3.2.2: An example of the ablation experiment. All the scores s are sorted in descending order. We mark the ground truth labels in red font, and the final classification results of each method in the red box. Source: Anonymous authors, <https://pixabay.com>. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). Modified by the author.

(mAP) that were often used in multi-label classification tasks [27].

Moreover, we conducted an ablation experiment. Specifically, we compared our proposed model with the image captioning model trained by real-world images. To investigate the contribution of using image captioning, we evaluated the effectiveness of the captioning model by BLEU score [14] that is commonly used in the image captioning task. In general, higher BLEU score means higher accuracy of the captioning results. To calculate BLEU score, we randomly chose 100 anime illustrations from Danbooru2020 dataset and respectively imported them into the captioning model trained by real-world images and that trained by the method in subsection 3.2.2.1.

3.2.3.4 Experimental Results

We show the comparison between the proposed method and other image classification methods in Table 5.1. By comparing with the baseline networks like ResNet-101 [16] which constructed the same latent space as ours, we can see the improvement of the performance after

the employment of GCN by the experimental results. Moreover, by comparing with the state-of-the-art methods based on GCN like ADD-GCN [27], we also confirm that introducing the domain-specific semantic features into the GCN model is effective.

We also show the ablation experimental results in Table 3.2.2. From this table, it can be confirmed that the higher the accuracy of the image captioning is, the higher the final classification performance is. Also, Fig. 5.4 shows the example of our method and the comparison method in the ablation experiments. As shown in this figure, the label *phone*, which is incorrectly classified as positive by the method using the captioning model trained for real-world image, is correctly classified as negative after using the captioning model trained for anime illustration. Therefore, the effectiveness of our method is verified.

3.2.4 Conclusions

We have proposed a GCN-based multi-modal multi-label attribute classification model in anime illustration using domain-specific semantic features. We capture the relationship between different attributes by introducing semantic features specific to anime illustration into GCN, and we further improve the classification accuracy. In addition, to generate more accurate image captioning for constructing our GCN model, we propose a trainable framework for the image captioning of anime illustration. The experiments show the effectiveness and the rationality of the proposed method.

Chapter 3.3

Latent Representation-based Semantic Understanding Method for Artificially Created Works Considering Label Hierarchy

3.3.1 Introduction

In recent years, the growth of the anime industry has advanced many studies on anime illustrations, such as illustration editing [53, 54], cartoon face generation [55, 56], and line-art colorization [57, 58]. Among these studies, the need for classification techniques for anime illustrations is also growing. The increasing number and complexity of anime illustrations have made it necessary for creators to have a way to efficiently manage and classify many illustrations. Classification techniques for anime illustrations help them automate the process of identifying and organizing specific elements within images, which can significantly streamline the animation production process. It is necessary to understand the contents of these illustrations to develop classification techniques for anime illustrations.

In the field of computer vision, image classification is an important task, which is the study of how computers can understand and analyze visual data from the world around us. Multi-label image classification is a variant of image classification in which an image can be labeled with multiple tags. This is in contrast to traditional image classification, which involves assigning a single label or class to an image. Multi-label image classification is useful in situations in which

an image may contain multiple objects or features relevant to the task at hand. For example, in medical image classification [44], an X-ray image may contain multiple organs or structures that need to be identified and labeled. Similarly, in recommendation systems [45], an image of a product may contain multiple features relevant to the user, such as color, size, and material.

In multi-label image classification, it is important to consider the relationships between labels to improve classification accuracy. This is because objects in images frequently co-occur, that is, they appear together in the same image. Graph convolutional networks (GCNs) [33] are a type of neural network that can be used to predict relationships between labels on graph structures. They are particularly useful in multi-label image classification because they can capture the dependencies between different labels and use this information to improve classification accuracy. For example, Chen et al. proposed a GCN-based multi-label image classification system that used a complete graph to model the correlation between various labels [26]. This approach was effective in improving classification accuracy. GCN-based approaches are expected to be particularly effective for anime illustrations because objects in anime illustrations frequently co-occur as well. Anime illustrations typically describe complex scenes with many elements, and GCNs can help us understand the relationships between these objects and improve classification accuracy.

However, classifying images of anime illustrations differs from classifying real-world images because anime illustrations are artificially produced and typically have distinct characteristics that may not be found in real-world images. For example, anime illustrations are frequently stylized and may describe objects or characters with exaggerated or unusual features. These features are known as attributes, and they can be important for accurately classifying anime illustrations. In addition to simple objects, it is necessary to consider the attributes of these objects when building classification methods for anime illustrations. For example, if we are classifying an image of an anime character, we may need to consider not only the object (e.g., character) but also the attributes of that object (e.g., hair color and eye shape). To the best of our knowledge, there has been no research specifically addressing the multi-label attribute classification task [21, 48] for anime illustrations. This is an open area of research that could be useful for applications such as image search and recommendation systems.

To use GCN-based multi-label classification techniques for anime illustrations to classify attributes, several problems need to be addressed.

Problem 1: Previous multi-label classification methods for anime illustrations [1, 46] have focused on capturing the relationships between attributes based on visual information about these illustrations. These methods typically involve training a classifier to recognize and classify different objects and attributes within an image. However, as mentioned above, anime illustrations are artificially created and frequently include more complex details than real-world images, making it difficult to accurately capture relationships between attributes using only visual information. To improve the accuracy of multi-label classification for anime illustrations, it is necessary to consider additional information or techniques that can better capture the relationships between attributes.

Problem 2: In previous GCN-based multi-label classification methods [26, 27], labels are typically treated equally when constructing graphs of co-occurrence relationships between them, which is reasonable because these labels represent "objects" rather than "attributes." However, there is a clear hierarchy between "attributes," i.e., upper and lower levels of inclusion. In other words, there are parent and child labels that have semantic relationships of subordination. Therefore, to improve the ability to capture the relationship between attributes, the hierarchy of attributes cannot be ignored.

In this study, we propose a hierarchical multi-label attribute classification of anime illustrations with GCNs to solve the above problems with the following considerations.

- To solve problem 1, we need to consider factors beyond visual information when classifying attributes in anime illustrations. One approach that has been proposed is to incorporate image captioning [50], which provides a more detailed representation of an image by using semantic information. However, current image-captioning models [34, 50] are trained only on real-world images, which can lead to less accurate classification results when applied to anime illustrations due to the differences between the two types of images, also known as a domain-shift problem. To mitigate this issue, we convert real-world images into anime-style images and pair them with the corresponding real-world image captions. This allows us to train an image-captioning model on a diverse set of images, including both real-world images and anime-style images, to make the feature spaces of the two domains more similar. In this way, the generation of captions for anime illustrations using only real-world images becomes feasible.

- To solve problem 2, we use a recursive regularization framework in the design of the loss function, as proposed in a previous study [59]. This approach assumes that if two labels are parent and child in a hierarchy, the classification from these two labels to other labels should be similar. In other words, a child-label classifier should inherit the characteristics of a parent-label classifier. Because the general anime illustration dataset excludes a defined attribute hierarchy, we use WordNet [60] to construct the attribute hierarchy in the anime illustration dataset based on the logical relationships between words. This allows us to incorporate the hierarchical structure of the labels into the classification process, and we can improve the overall accuracy of the model.

In summary, the contribution of this study can be highlighted as follows. First, we propose a GCN-based model for the multi-label attribute classification of anime illustrations that combines both visual and semantic features (i.e., captioning features) suitable for anime illustrations. Second, to improve the classification accuracy, we incorporate the regularization term influenced by label hierarchical relationships into the loss function because attributes in anime illustrations have hierarchical relationships. Finally, we construct a hierarchical structure of attributes in the anime illustration dataset based on the defined logical relationships between words, which helps better reflect the relationships between different attributes in the classification process.

3.3.2 Proposed Method

This section outlines the proposed model for multi-modal multi-label attribute classification of anime illustrations using GCNs. In 3.3.2.1, we detail the construction of a new captioning framework for anime illustrations. In 3.3.2.2, we describe the general flow of the GCN-based classification process. Finally, in 3.3.2.3, we explain the final classification and the loss function used to train the model. Fig. 3.3.1 shows an overview of the proposed model.

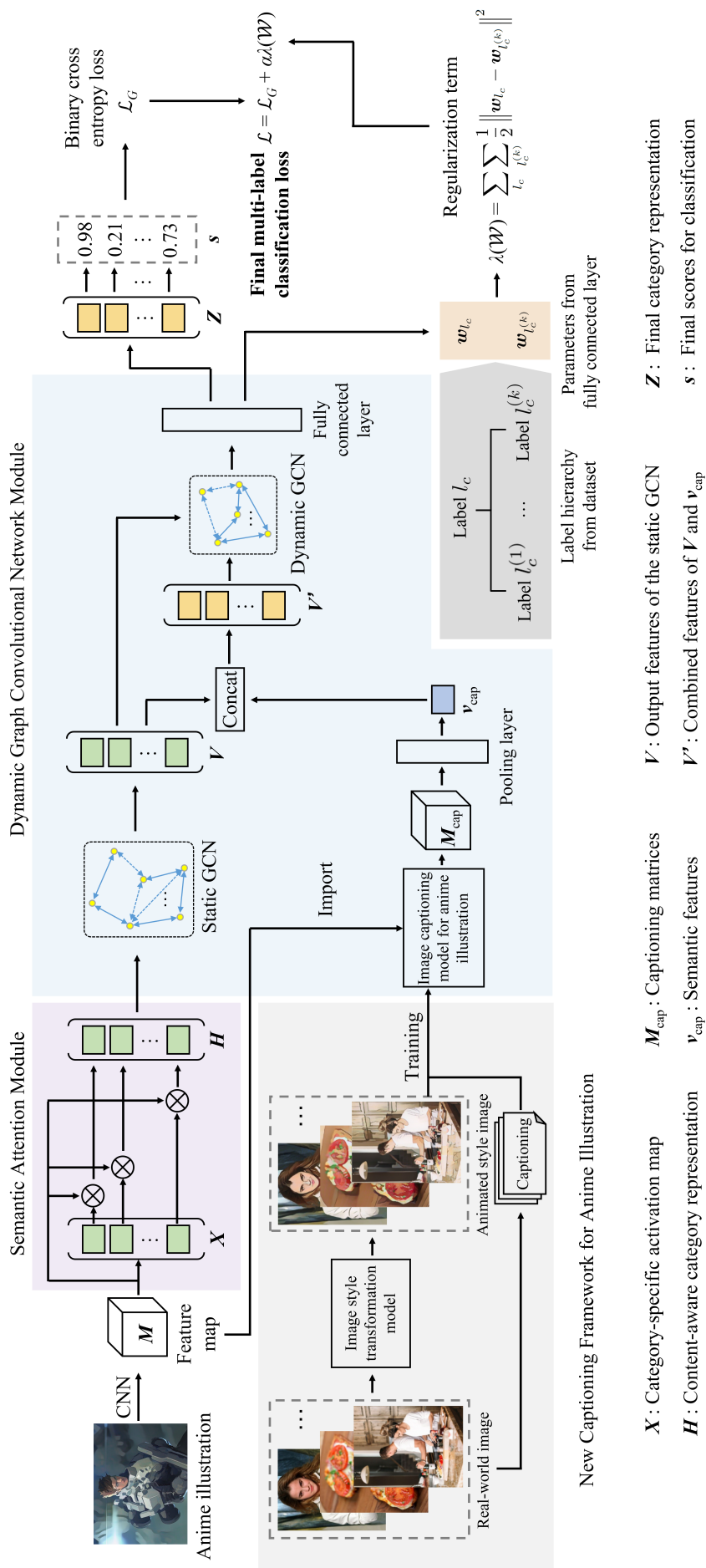


Figure 3.3.1: Overview of the proposed model for hierarchical multi-label attribute classification of anime illustrations. Source: Anonymous authors, <https://pixabay.com>. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). Modified by the author.

3.3.2.1 New Captioning Framework for Anime Illustrations

In this subsection, we demonstrate the process of generating features that represent the semantic characteristics of anime illustrations.

First, we use an image style transformation model, specifically the white-box cartoonization model [51], to transform real-world images into animated style images. The white-box cartoonization model works by using a CNN to analyze input images and to identify features such as edges, shapes, and textures. These features are then used to generate a cartoon version of these input images, allowing us to obtain high-quality sets of real-world and animated style images.

Next, we feed the real-world images into a pretrained captioning model [34] to obtain captioning results. In most cases, the caption of a real-world image can be considered the actual caption of the corresponding animated style image because the transformation process does not significantly alter the content of the image. To reduce the gap between captioning results for animated style and real-world images, we train the captioning model using both the generated captions of corresponding real-world images and the animated style images. This process helps resolve the domain-shift issue, which is further discussed in Section 3.3.3.

3.3.2.2 GCN-based Multi-label Attribute Classification

This subsection outlines the general flow of GCN-based multi-label attribute classification. The model, inspired by a uni-modal version of a prior GCN [27], consists of two main components, as depicted in Fig. 3.3.1: the semantic attention module and the dynamic GCN module.

In the semantic attention module, we first use a CNN backbone to extract convolutional feature maps $\mathbf{M} \in \mathbb{R}^{H \times W \times D}$ from the input image, where H denotes the height of $\mathbf{M}$, W denotes the width of $\mathbf{M}$, and D denotes the number of channels of $\mathbf{M}$. Then, we generate a set of category-specific activation maps $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ using class activation mapping [35], where C represents the number of categories. Then, we convert $\mathbf{M}$ into a content-aware category representation $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_C]^\top \in \mathbb{R}^{C \times D}$ using the activation maps $\mathbf{X}$ as follows:

$$\mathbf{h}_c = \sum_{i=1}^H \sum_{j=1}^W x_{i,j}^c \mathbf{m}_{i,j}, \quad (3.3.1)$$

where $x_{i,j}^c$ denotes the (i, j) -th weight of c -th activation map $\mathbf{X}_c \in \mathbb{R}^{H \times W}$, and $\mathbf{m}_{i,j} \in \mathbb{R}^D$ denotes

the (i, j) -th feature vector of the feature map $\mathbf{M}$.

In the dynamic GCN module, we import $\mathbf{H}$ sequentially into a static GCN and a dynamic GCN to obtain different representations of the label relations for specific input image, and also in this module, the semantic features from the anime illustrations obtained by the captioning model are introduced.

We first feed $\mathbf{H}$ into a single-layer static GCN, defined as $\mathbf{V} = LReLU(\mathbf{A}_{st}\mathbf{H}\mathbf{W}_{st}) \in \mathbb{R}^{C \times D}$, where $\mathbf{A}_{st}$ represents the correlation matrix, and $\mathbf{W}_{st}$ represents the state update weights. $LReLU(\cdot)$ denotes the activation function LeakyReLU [36], which is a variant of the standard ReLU function that allows a small amount of negative values to pass through. This is useful for preventing the model from getting stuck in a state where all the neurons are deactivated, which can occur with standard ReLU. After that, we use the image-captioning model trained for anime illustrations to obtain the captioning matrixes $\mathbf{M}_{cap} \in \mathbb{R}^{H \times W \times D}$. These matrixes are then transformed into semantic features $\mathbf{v}_{cap} \in \mathbb{R}^D$ that can be integrated into the multi-label classification model using a pooling layer. This process is defined as follows:

$$\mathbf{v}_{cap} = p(\mathbf{M}_{cap}), \quad \text{s.t. } \mathbf{M}_{cap} = cap(\mathbf{M}), \quad (3.3.2)$$

where $p(\cdot)$ denotes the pooling layer, and $cap(\cdot)$ denotes the captioning model from the previous subsection. To improve the classification performance by incorporating semantic information from the captions, we combine the output features $\mathbf{V}$ from the static GCN with the semantic features $\mathbf{v}_{cap}$ obtained from the captioning model. This is achieved by concatenating the two feature vectors: $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_C]^\top \in \mathbb{R}^{C \times D}$ and $\mathbf{V}_{cap} = [\mathbf{v}_{cap}, \mathbf{v}_{cap}, \dots, \mathbf{v}_{cap}]^\top \in \mathbb{R}^{C \times D}$. The resulting combined feature is $\mathbf{V}' = [\mathbf{V}, \mathbf{V}_{cap}] \in \mathbb{R}^{C \times 2D}$, which is then input into the dynamic GCN. The output of the dynamic GCN, denoted as $\mathbf{Z} \in \mathbb{R}^{C \times D}$, is calculated using the following equation:

$$\mathbf{Z} = LReLU(\mathbf{A}_{dy}\mathbf{V}\mathbf{W}_{dy}), \quad (3.3.3)$$

where $LReLU(\cdot)$ denotes the LeakyReLU activation function, $\mathbf{A}_{dy}$ denotes the dynamic correlation matrix, and $\mathbf{W}_{dy}$ denote the state update weights. $\mathbf{A}_{dy}$ is calculated using a *conv* layer with weights $\mathbf{W}_A$ applied to $\mathbf{V}'$, followed by the sigmoid activation function $\sigma(\cdot)$ as $\mathbf{A}_{dy} = \sigma(\mathbf{W}_A \mathbf{V}')$. As a result, this GCN flow can capture the relationships between different attributes in images

while also incorporating textual descriptions of anime illustrations using a captioning model, which can improve classification accuracy.

3.3.2.3 Final Classification and Loss

In this subsection, we describe the process of the final classification and the calculation of the loss function. The output of the dynamic GCN, $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_C]^\top$, is used for the final classification. Specifically, we input each vector $\mathbf{z}_c$ of the final category representation $\mathbf{Z}$ into a fully connected layer to obtain the predicted scores s^c for category c . These scores are then concatenated to form the final score vector $\mathbf{s} = [s^1, s^2, \dots, s^C]^\top$. According to previous works [26–28], the loss function $\mathcal{L}_G(\mathbf{y}, \mathbf{s})$ can be defined as follows:

$$\mathcal{L}_G(\mathbf{y}, \mathbf{s}) = \sum_{c=1}^C y^c \log(\sigma(s^c)) + (1 - y^c) \log(1 - \sigma(s^c)), \quad (3.3.4)$$

where $\mathbf{y} \in \mathbb{R}^C$ represents the ground truth labels for an image, and $y^c = \{0, 1\}$ indicates whether label c is present or absent in the image.

In a general multi-label classification task, it is common to train the entire model using a loss function $\mathcal{L}_G$. This loss function is used to measure the difference between the predicted labels and their corresponding true labels, and it is used to update the parameters of the model to minimize this difference. However, when classifying the attributes of anime illustrations, it is necessary to consider the hierarchical relationship among labels. In a hierarchical structure, parent labels are the upper topics of child labels, and when two labels are relatively close hierarchically, they tend to have similar parameters due to their hierarchical relationship. For example, in the hierarchical structure depicted in Fig. 3.3.2, the labels *hair type* and *twintails* are likely to have similar parameters because there is an edge connecting them, which means *twintails* is a child label of *hair type*, and it is likely to inherit some of the characteristics of its parent label.

Therefore, we apply recursive regularization over the fully connected layer of the dynamic GCN, similar to the approach described in [61]. Specifically, we denote the parameters in the final fully connected layer for label l_c in the hierarchy as $\mathbf{w}_{l_c}$, and the set of all such parameters

as $\mathcal{W} = \{\mathbf{w}_{l_c}\}$. We regularize these parameters using the following regularization term:

$$\lambda(\mathcal{W}) = \sum_{l_c} \sum_{l_c^{(k)}} \frac{1}{2} \|\mathbf{w}_{l_c} - \mathbf{w}_{l_c^{(k)}}\|^2, \quad (3.3.5)$$

where $l_c^{(k)}$ refers to a child of l_c . The final loss function $\mathcal{L}$ is then calculated as follows:

$$\mathcal{L} = \mathcal{L}_G + \alpha \lambda(\mathcal{W}), \quad (3.3.6)$$

where α denotes the penalty parameter.

In this way, the hierarchy of attributes ignored in previous studies is considered in the calculation of the loss function, which can improve the ability to capture the relationships between attributes.

3.3.3 Experimental Results

In this section, we present experimental results to demonstrate the effectiveness of our proposed method. In 3.3.3.1, we detail the training of our captioning model for anime illustrations. In 3.3.3.2, we introduce the anime illustration dataset used in the experiment and explain how we constructed the hierarchical structure for the labels in this dataset. In 3.3.3.3, we describe the experimental conditions and introduce the comparison methods and evaluation metrics used in this experiment. Finally, in 3.3.3.4, we present the experimental results.

3.3.3.1 Training Settings of Captioning Model for Anime Illustrations

To train the captioning model for anime illustrations, we first generated a dataset of animated style images by applying the image style transformation model to real-world images. To do this, we used the Flickr8k dataset [52]. The Flickr8k dataset is a dataset of 8000 real-world images and their associated captions, and it is commonly used for training and evaluating image captioning models. Each of the images in this dataset has been annotated with five different captions describing the contents of the image. We imported the images from this dataset into the image style transformation model [51] mentioned in III-A to generate their corresponding animated style images. After that, to train the captioning model, we used the generated animated style images as input and the captions of their corresponding real-world images as the ground

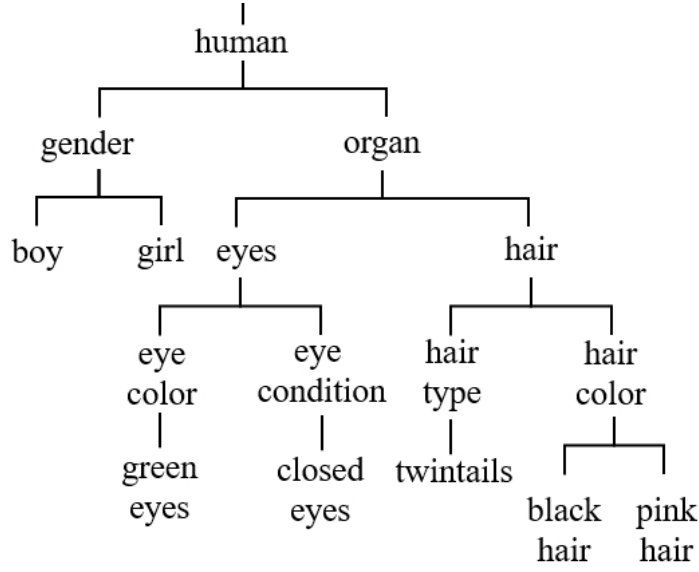


Figure 3.3.2: Part of the label hierarchy of anime illustration dataset [1].

truth. The goal was to train the model to accurately generate captions for the generated animated style images based on the captions provided for the real-world images. Specifically, we followed a previous study [34] to train the model. Once the model was trained, we could input an anime illustration and generate its caption based on the information from the training data. This allowed us to effectively capture the semantic information of anime illustrations.

3.3.3.2 Anime Illustration Dataset and Construction of Label Hierarchy

In this experiment, we used the Danbooru 2020 dataset [1] to train the multi-label attribute classification model. The Danbooru 2020 dataset is a comprehensive anime illustration dataset with over 4.2 million illustrations and 130 million labels. To construct accurate and convincing hierarchical relations for the labels in this dataset, we used the semantic relations of words defined in a large dictionary, WordNet [60], where words are grouped into a hierarchy defined by superordinate and subordinate relations. To select labels from the Danbooru 2020 dataset to be used for classification, we first filtered out the 500 most frequently appearing labels. Because some of these 500 labels are difficult to classify into a semantic hierarchy, we need to remove them and use the remaining tags to build the hierarchy. In our experiment, the following labels are removed.

- Labels of the title of the work, the name of the character, the name of the illustrator, etc. (e.g., *hatsune miku*)
- Labels describing information about the illustration, not the content of the illustration (e.g., *absurdres*)
- Labels describing the art style to which the illustration belongs (e.g., *traditional media*, *monochrome*, *sketch*)
- Labels describing the character’s facial expression, movement, or pose (e.g., *happy*, *standing*)
- Labels describing the layout of the illustration (e.g., *upper body*, *cowboy shot*)

After removing the above labels, the remaining 359 labels were matched with words in WordNet [60] to construct the hierarchical relationship. Finally, seven layers of the hierarchical structure were constructed. Figure 3.3.2 shows a part of this hierarchy. As depicted in the figure, the directly connected parent and child nodes are labels with an inclusive relationship. Next, we randomly selected 25,000 anime illustrations from the dataset containing 359 labels in the hierarchy, of which 75% were used as the training set and 25% as the validation set.

3.3.3.3 Experimental Conditions, Comparison Methods, and Evaluation Metrics

We used ResNet-101 [16] as the backbone of the GCN-based attribute classification model. The negative slope of the LeakyReLU activation function, which was used in the dynamic GCN module, was set to 0.2. To improve the generalization of the model, we used data augmentation techniques on the input images. We randomly cropped the images, resized them to 448×448 pixels, and horizontally flipped them, which can artificially increase the size of the training set by creating new images from the existing ones and help the model learn to be more robust to variations in the input data. To optimize the model, we used stochastic gradient descent as the optimizer. We set the momentum decay to 0.9 and the weight decay to 1.0×10^{-4} . The learning rates for the different modules of the model were initially set at 0.5 for the semantic attention module and the dynamic GCN module, and 0.05 for the backbone CNN.

Table 3.3.1: Performance comparison between our model and other image classification models. We mark the best results as bold.

Method	OP	OR	OF1	CP	CR	CF1	mAP
Resnet-101 [16]	61.0	56.5	58.7	60.4	55.2	57.7	60.4
SSGRL [25]	69.0	57.2	62.5	70.2	58.2	63.6	68.6
DAN [21]	64.9	51.0	57.1	66.5	56.8	61.3	64.6
ML-GCN [26]	63.8	60.5	62.1	60.1	54.2	57.0	62.3
ADD-GCN [27]	69.6	64.1	66.7	66.4	60.2	63.1	68.1
P-GCN [28]	69.1	58.9	63.6	72.8	56.4	63.6	70.2
Ours	73.1	66.8	69.8	69.9	59.1	64.0	71.1

To evaluate the effectiveness of the proposed method for multi-label attribute classification of anime illustrations, we compared it with several other methods. These methods were as follows: 1) ResNet-101 [16], an extensively used CNN for image classification tasks; 2) DAN [21], a method that uses CNNs to learn discriminative features for multi-label attribute classification on real-world images; 3) SSGRL [25]; a method that uses CNNs and a graph propagation mechanism to improve multi-label classification performance; 4) ML-GCN [26], a method that uses GCNs to model the correlations between labels in a multi-label classification task; 5) ADD-GCN [27], a state-of-the-art method that constructs dynamic diagrams to describe label relationships in images to improve multi-label classification performance; 6) P-GCN [28], another state-of-the-art method that uses a GCN to improve multi-label classification performance, which is an extended version of ML-GCN. The proposed method used ResNet-101 as its backbone, and these methods were trained using similar hyperparameters for a fair comparison. We evaluated the performance of each method on multi-label attribute classification for anime illustrations and compared their results with the proposed method to verify its effectiveness.

We adopted the following evaluation metrics from previous studies [26–28]: Averages of overall precision (OP), recall (OR), and F1 score (OF1); averages of per-class precision (CP), recall (CR), and F1 score (CF1). Furthermore, we used the average precision (AP) and the mean average precision (mAP), which are commonly used metrics for multi-label classification tasks [27]. Generally, the average overall F1 (OF1), average per-class F1 (CF1), and mAP are considered the most important metrics for evaluating performance.

In addition to the main experimental evaluation, we also conducted an ablation study to further investigate the effectiveness of our proposed method. In the ablation study, we controlled two key variables as follows:

Table 3.3.2: Ablation experimental results. In the ablation experiment, we randomly selected 100 anime illustrations with ground truth captions from the Danbooru 2020 dataset and calculated BLEU score [14] and mAP.

Method	Captioning model used in the method		Loss function	mAP
	Details	BLEU [14]		
CM1	Trained for real-word image	44.5	$\mathcal{L}_G$	64.2
CM2			$\mathcal{L}_G + \alpha\lambda(\mathcal{W})$	67.1
CM3	Trained for anime illustration	55.2	$\mathcal{L}_G$	66.8
Ours			$\mathcal{L}_G + \alpha\lambda(\mathcal{W})$	70.3

- **The captioning model used in the method.** We compared our proposed method with different variations of it, by using different pre-trained captioning models: trained by real-world images, or trained for anime illustrations. This allowed us to investigate the effect of the captioning model on the overall performance of the model.
- **The regularization term influenced by the label hierarchy.** We compared our proposed method with another method that does not consider the regularization term influenced by the label hierarchy when calculating the loss function. This allowed us to understand the effect of the hierarchical structure on the final classification.

Specifically, we used three comparison methods (CM1, CM2, and CM3), and the details are presented in Table 3.3.2. To evaluate the effectiveness of the captioning model, the BLEU score [14] was used. The BLEU score is a commonly used measure of the quality of generated captions in image-captioning tasks, with higher scores indicating better results. To obtain the BLEU scores, 100 anime illustrations from the Danbooru 2020 dataset were randomly selected and fed into the captioning models trained on real-world images and the model trained using the method described in III-A, respectively. This allowed us to compare the performance of the different methods and evaluate the effectiveness of the captioning model in the proposed method.

3.3.3.4 Experimental Results and Discussions

Table 3.3.1 compares the proposed method with other image classification methods. The experimental results show that our GCN outperforms baseline networks such as ResNet-101 [16], which create the same latent space as ours. We also discovered that introducing semantic features and a label hierarchy to the GCN-based model improves performance compared with the

state-of-the-art GCN-based methods such as ADD-GCN [27].

In addition, in Table 3.3.2, we present the ablation experimental results. Comparing the performance of CM1, CM2, and our method, the higher the accuracy of image captioning, the higher the final classification performance. Furthermore, comparing the performance of CM1, CM3, and our method, we can confirm that introducing the regularization term influenced by the label hierarchy when calculating the loss function has a positive effect on the final classification performance.

We also perform qualitative evaluations to demonstrate the effectiveness of our method. For the effect of introducing the captioning model on the final classification performance, we demonstrate some examples in the ablation experiments in Fig. 3.3.3. Specifically, to control variables, we compare the performance of our method with that of CM2 used as a comparison method in the ablation experiment because both methods differ only in the captioning model used in the model. In Fig. 3.3.3, we present the captioning results generated by the captioning models used in these methods and the final classification results of the two methods using histograms. As depicted in the figure, although the captioning results generated by the captioning model for anime illustration are not always completely correct, the proposed method still outperforms CM2, in terms of not only the successful classification of correct labels but also the exclusion of incorrect labels. We also evaluate the effect of introducing the hierarchical structure on the performance of the model. Similarly, we adopt CM3, which uses the same captioning model as the proposed method but does not introduce a hierarchical structure when calculating the loss function, as a comparison method. In the example presented in Fig. 3.3.4, the predicted scores of the labels are increased after the introduction of the hierarchical structure. Specifically, the label *black hair*, which is incorrectly classified as negative by CM3, is correctly classified as positive after introducing the hierarchical structure because of the increase in the score of *long hair*, which has a close relationship with *black hair* in the hierarchical structure. Therefore, the effectiveness of our method is verified.

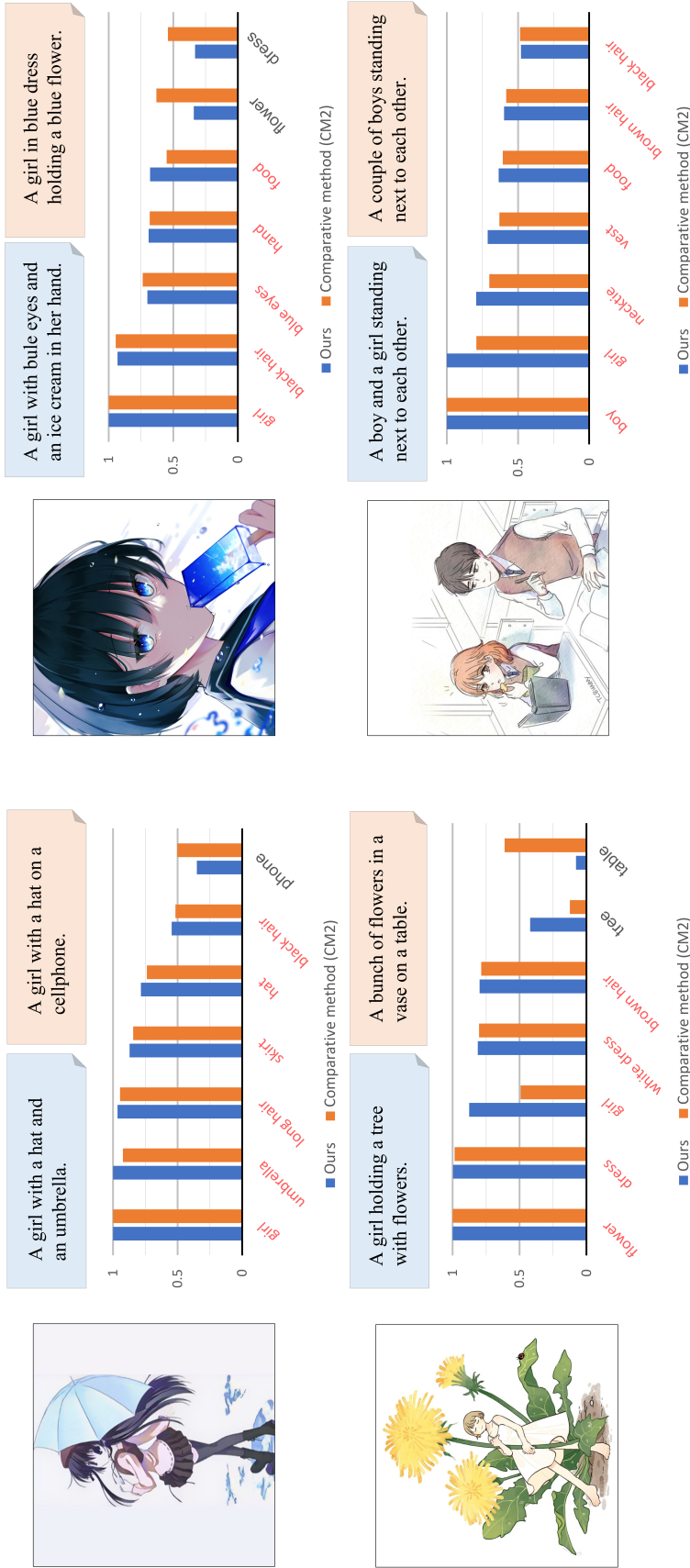
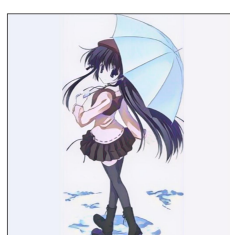
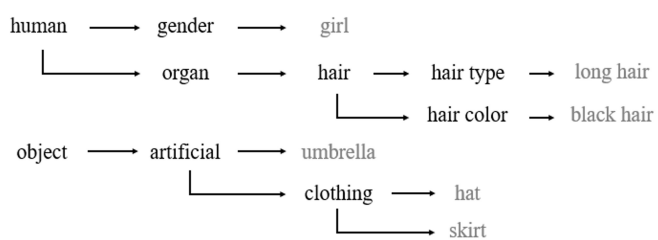


Figure 3.3.3: Examples of the effect of the captioning model on the final classification in the ablation experiments. The captioning results of each illustration generated by the captioning models used in the proposed method and CM2 are shown in the boxes of blue and orange backgrounds, respectively. We also show the predicted scores of the labels using histograms. Labels are considered positive if the predicted score is greater than 0.5. The horizontal axis labels of the histograms are the concatenated sets of the top-7 highest scoring labels predicted by the proposed method and CM2, and the ground truth labels are marked in red font. Source: Anonymous authors, <https://pixabay.com>. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). Modified by the author.



Anime Illustration



Hierarchical structure of labels

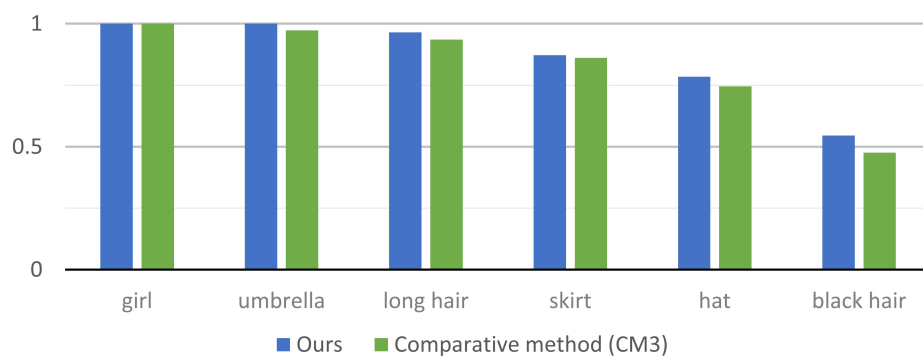


Figure 3.3.4: An example of the effect of the hierarchical structure on the final classification in the ablation experiments. Source: Anonymous authors, <https://pixabay.com>. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). Modified by the author.

3.3.4 Conclusion

In this study, we developed a GCN-based hierarchical multi-label attribute classification model for anime illustrations. By introducing semantic features specific to anime illustrations into the GCN, we could capture the relationships between different attributes and improve classification accuracy. To generate more accurate image captions for our GCN-based model, we proposed a trainable framework for the image captioning of anime illustrations. Furthermore, we constructed a hierarchical structure of attributes in an anime illustration dataset and used this hierarchy to further capture relationships between attributes. Our experiments demonstrated the effectiveness and rationality of the proposed method.

Chapter 4

Adversarial Perturbation Based Continual Personalization Diffusion Model for Mitigating Concept Bleed-through

4.1 Introduction

Recent studies have investigated the use of diffusion models [62] to generate high-quality images. Text-to-image diffusion models, e.g., the DALL-E [63] and stable diffusion [19] models, have become the mainstream in research on text-to-image models, and these models are currently used for various tasks, including image generation from arbitrary text, image editing [64, 65], super-resolution [66], and inpainting [67, 68]. Despite the extensive training of diffusion models, they may struggle to generate unique or personalized concepts that are not present in the training corpus, e.g., personalized styles or specific faces. There has been an increasing trend toward implementing personalization methods in text-to-image diffusion models, including textural inversion [10], DreamBooth [7], and LoRA [22]. In addition to these methods, which are employed to realize personalization to a single concept, custom diffusion methods [30] learn multiple concepts. These methods enable the creation of images related to a new specific concept (an object or style) based on a few reference examples, and the learning process is efficient by changing only a small portion of the weights in the entire pipeline of the diffusion model, which results in both fast concept acquisition and lightweight model updates.

These methods have demonstrated impressive results in generating personalized content; however, they suffer from a critical limitation when applied to the continual learning context, which is a common use case for personalized generative models. In continual learning, the models must learn multiple concepts sequentially over time without previous training data. This paradigm is essential to reduce computational cost and preserve data privacy because storing and retraining on previous data may be infeasible. However, existing methods for personalized diffusion models struggle with catastrophic forgetting, where newly learned concepts overwrite previously acquired ones, and catastrophic-neglect, where images generated by the models miss key concepts mentioned in the prompt. To address these issues, recent studies have introduced methods, e.g., C-LoRA [23] and CIDM [24], which enhance the continual learning for diffusion models. These methods have demonstrated promising results; however, they still face a fundamental problem, which we call concept bleed-through [23]. As shown in Fig. 4.1, concept bleed-through occurs when learning a new concept inadvertently alters or distorts the generation of similar previously learned concepts, thereby leading to unintended hybrid or incorrect outputs.

There have been attempts to address concept bleed-through at the model structure level. However, an effective solution to prevent concept bleed-through while maintaining high-quality personalized image generation remains an open problem. In addition, these structural modifications are not directly applicable to existing personalization frameworks and frequently introduce additional model complexity, which limits their practicality in real-world applications. Thus, we focus on preprocessing the training images.

In the continual learning of diffusion models, handling similar concepts is challenging because the semantic features contained in images are inherently similar, which makes it difficult to avoid mapping them to closely positioned features in the latent space. Thus, we consider methods to manipulate the semantic information of images while preserving their visual information as much as possible. Inspired by recent advances in adversarial attacks on diffusion models [69], we propose a method to mitigate the concept bleed-through problem in personalized diffusion models. Adversarial attacks on diffusion models typically involve the introduction of carefully crafted perturbations to the input images to alter their semantic representations in a manner that confounds the model while remaining imperceptible to human observers. These adversarial attacks exploit the vulnerabilities of the diffusion models by modifying the latent representations

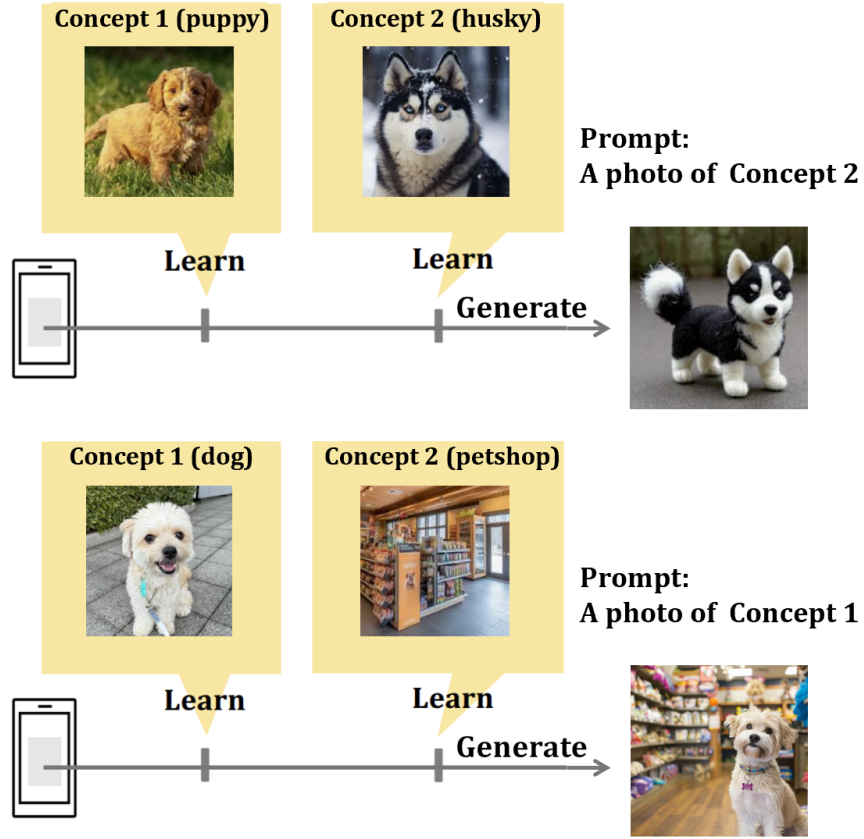


Figure 4.1: Examples of the concept bleed-through problem in the continual learning of personalized generative models. When handling nearby concepts (e.g., *puppy* and *husky*) or related concepts (e.g., *dog* and *petshop*), although the prompt only involves a single learned concept, the generated image is clearly influenced by the other learned concepts.

of the training data, thereby leading to significant deviations in the generated outputs.

In this chapter, we propose a training method for personalized diffusion models using adversarial perturbation. Specifically, we inject subtle adversarial perturbations into the training images used to personalize the diffusion model. These perturbations are designed to shift the semantic representation of the concepts in the latent space, which ensures that newly learned concepts occupy distinct regions that do not interfere with the previously learned concepts. By adopting this strategy, the proposed method effectively prevents concept bleed-through while preserving the fidelity and consistency of personalized image generation.

The technical contributions of this study are summarized as follows.

- We introduce a training strategy based on adversarial perturbation for personalized diffusion models, where adversarial perturbations are leveraged to prevent concept bleed-

through in continual learning scenarios. To the best of our knowledge, this is the first study to introduce adversarial perturbations into the continual learning task of diffusion models.

- The proposed method systematically shifts the semantic representation of personalized concepts within the latent space, which ensures that the newly learned concepts do not interfere with previously acquired concepts, thereby enhancing the ability of the diffusion models to learn new concepts over time without catastrophic forgetting or unintended concept mixing.
- The proposed method operates at the data level rather than modifying the structure of the model, which ensures broad compatibility with existing personalization frameworks while avoiding additional model complexity.

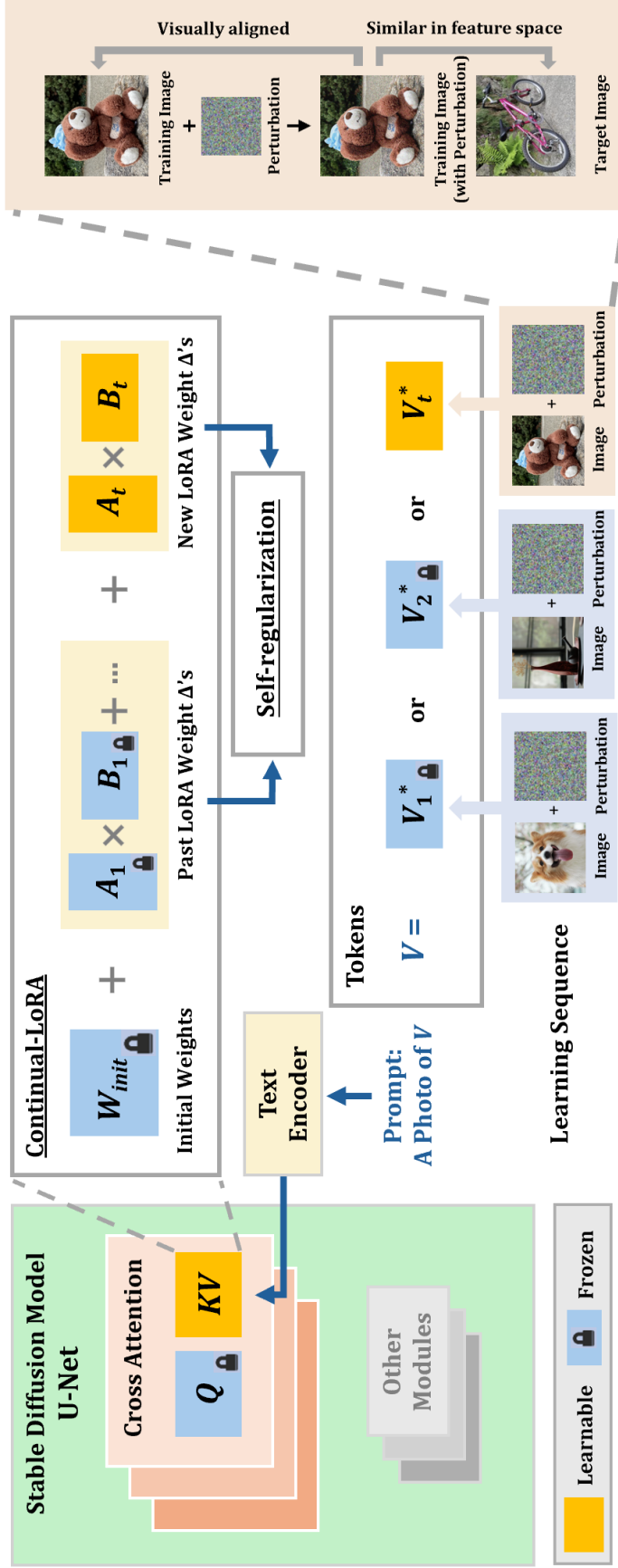


Figure 4.2: Overview of the proposed method. We update the key-value (K - V) projection in the U-Net cross-attention modules of stable diffusion using a continual, self-regulating low-rank weight adaptation. In addition, we add adversarial perturbations to the training data to achieve effective concept separation in the latent space.

4.2 Methodology

In this section, we introduce the proposed method in detail. An overview of the proposed method is shown in Fig. 4.2. It is worth clarifying that the first two subsections serve to explain the baseline continual learning framework in a simple diffusion model. The key novelty lies in last subsection, where we introduce our adversarial perturbation strategy, and our aim is to isolate the contribution of adversarial perturbation under a simple continual learning framework.

4.2.1 Parameter Modification in Continual LoRA

Inspired by [23], and considering that the current mainstream continuous learning frameworks for personalized image generation are mostly based on LoRA, we adopt a LoRA-based fine-tuning method and introduce adversarial perturbations to the training images to shift their semantic representations within the latent space. This ensures that the newly learned concepts do not interfere with the previous concepts, thereby preventing the concept bleed-through. We use LoRA due to its efficiency in updating only a small fraction of parameters while preserving most of the pretrained weights. Unlike full fine-tuning methods, which adjust all parameters, LoRA-based adaptation is considerably more memory-efficient and can be applied sequentially without increasing the size of the model significantly. Thus, it is well-suited for continual learning scenarios where the model updates must be performed iteratively without catastrophic forgetting.

In the continual diffusion setting, we learn N personalization tasks $t \in \{1, 2, \dots, N-1, N\}$, where N denotes the total number of concepts that will be learned by the model. Following the results of the custom diffusion process [30], we modify a small subset of parameters in the stable diffusion model [19], specifically the cross-attention layers in U-Net because the cross-attention parameters of U-Net are most sensitive to changes during the personalization of the diffusion model. Here, we freeze the query features Q and parameterize the weight updates $W^{K,V} \in \mathbb{R}^{D_1 \times D_2}$ of the key features K and the value features V using LoRA [22]:

$$\begin{aligned} W_t^{K,V} &= W_{t-1}^{K,V} + A_t^{K,V} B_t^{K,V} \\ &= W_{init}^{K,V} + \left[\sum_{t'=1}^{t-1} A_{t'}^{K,V} B_{t'}^{K,V} \right] + A_t^{K,V} B_t^{K,V}, \end{aligned} \quad (4.1)$$

where $A_t^{K,V} \in \mathbb{R}^{D_1 \times r}$ and $B_t^{K,V} \in \mathbb{R}^{r \times D_2}$ correspond to the key and value projection matrices at

time step t , which encode the learned concept information. Note that r is a low-rank hyperparameter, and D_1 and D_2 denote the input and output dimensions of the key and value projection matrices in the cross-attention layers, respectively. Furthermore, $W_{init}^{K,V}$ is the initial values from the pre-trained model. The above approach allows for efficient parameter adaptation while maintaining previous knowledge.

4.2.2 Self-regularization

To mitigate catastrophic forgetting and preserve the previously learned concepts in personalized generative models, we introduce a self-regularization loss that constrains the parameter updates. This regularization term shown in the following equation ensures that updates affecting new concepts do not interfere destructively with the existing knowledge:

$$\mathcal{L}_{\text{forget}} = \left\| \sum_{t'=1}^{t-1} A_{t'}^{K,V} B_{t'}^{K,V} \right\| \odot A_t^{K,V} B_t^{K,V} \Big\|_F^2, \quad (4.2)$$

where $\odot$ represents element-wise multiplication, and $\|\cdot\|_F^2$ denotes the squared Frobenius norm. This self-regularization loss penalizes overlapping activations between the previously learned (t') and the newly introduced (t) concepts in the latent space, which effectively reduces semantic interference.

By integrating the self-regularization into the training objective, the proposed method achieves an effective balance between continual adaptation and knowledge preservation, which enables the personalized diffusion model to learn a sequence of fine-grained concepts, thereby resulting in more stable image generation. Specifically, when learning the personalization task t , we minimize the following total loss:

$$\min_{W_t^{K,V} \in \theta} \mathcal{L}_{\text{diff}}(i, \theta) + \lambda \mathcal{L}_{\text{forget}}, \quad (4.3)$$

where i is the input training data of the new concept, $\mathcal{L}_{\text{diff}}$ is the loss function for the stable diffusion model θ [19], and λ is a hyperparameter selected with a simple exponential sweep.

4.2.3 Training on Images with Adversarial Perturbations

The self-regularization enhances the model’s ability to resist catastrophic forgetting, ensuring that previously learned concepts in the learning sequence are retained; however, it does not effectively mitigate the adverse interactions between different concepts, particularly in cases where semantically similar concepts are introduced sequentially. As mentioned in the introduction, existing continual learning methods for generative models, e.g., C-LoRA and STAMINA, primarily focus on mitigating forgetting but do not explicitly address concept interference. To overcome this limitation, we introduce adversarial perturbations to the training images, which strategically shift the semantic features in the latent space while preserving the quality and visual consistency of the image. From a theoretical perspective, adversarial perturbations are designed to maximize the divergence between representations in the latent space by targeting directions of high sensitivity in the model’s gradient. Specifically, the optimization encourages the perturbed inputs to move away from the existing concept’s latent features, thereby enforcing semantic separation. Mathematically, this aligns with maximizing the cosine distance, akin to contrastive representation learning.

Inspired by previous studies [69], for an image x_t , we optimize the perturbation δ using targeted PGD [70]:

$$\begin{aligned}\delta &:= \arg \min_{\|\delta\| \leq \epsilon} \mathcal{L}_{\mathcal{E}}(x_t, \delta, x_t^{\text{pr}}) \\ &= \arg \min_{\|\delta\| \leq \epsilon} \left\| \mathcal{E}(x_t^{\text{pr}}) - \mathcal{E}(x_t + \delta) \right\|_2^2,\end{aligned}\tag{4.4}$$

where ϵ is the perturbation budget, $\mathcal{E}$ and x_t^{pr} are the encoder of the diffusion model and the given target image for x_t , respectively. Note that the selection of x_t^{pr} is critical in terms of ensuring effective perturbation-induced separation in the latent space. Specifically, for each input image x_t associated with a new concept, we sample x_t^{pr} from a pool of images that are semantically dissimilar to $x_1, x_2, \dots, x_{t-1}$, based on their precomputed latent representations. This ensures that the perturbation δ shifts the latent encoding of x_t away from the region associated with $x_1, x_2, \dots, x_{t-1}$. As a result, we promote better separability between the learned concepts by encouraging the features of different concepts to occupy distinct and distant regions of the latent space.

By minimizing the loss $\mathcal{L}_{\mathcal{E}}$ in Eq. (4.4), the perturbed image $x_t + \delta$ is embedded into the latent

space aligned with the target image, which effectively enforces a shift in the semantic representation. This adjustment prevents undesired overlaps between the newly learned concepts and the previously stored knowledge. By constraining the perturbation budget ϵ during optimization, the perturbation is kept below a certain threshold, preventing degradation in image quality or unintended changes in image semantics. Consequently, the proposed method significantly reduces the risk of concept interference, which ensures that the personalized concepts remain distinct over time. The proposed method is novel because we explicitly manipulate the latent space via adversarial perturbations, thereby creating a more robust framework for continual learning in generative models.

To further mitigate the interference between concepts, we follow a previously proposed custom tokenization strategy [2, 23]. Here, we define N personalized tokens $V_1^*, V_2^*, \dots, V_N^*$, which are initialized randomly rather than using less frequent words. In addition, to avoid interference with existing concepts, the object names are removed from the input sequence. During inference, the learned embeddings replace these tokens to generate images of multiple learned concepts simultaneously. This tokenization strategy ensures that different concepts retain unique representations in both the textual and latent spaces, which reduces the likelihood of semantic entanglement.

It is worth noting that our method essentially establishes a connection between the customized text embeddings and the slightly misaligned image features through targeted guidance using the target image. Therefore, it does not compromise the model’s original ability to generate the concepts present in the target image.

4.3 Experiments

4.3.1 Main Experimental Settings

To evaluate the effectiveness of the proposed adversarial perturbation-based training strategy, we performed experiments on personalized diffusion models under a continual learning setting. Here, the goal was to assess whether the proposed method mitigates concept bleed-through while maintaining high-quality continual personalized image generation. The quantity and content of the experiments followed those of [23] and [24].

Implementation Details: Following previous studies that utilized existing methods [23, 30], we

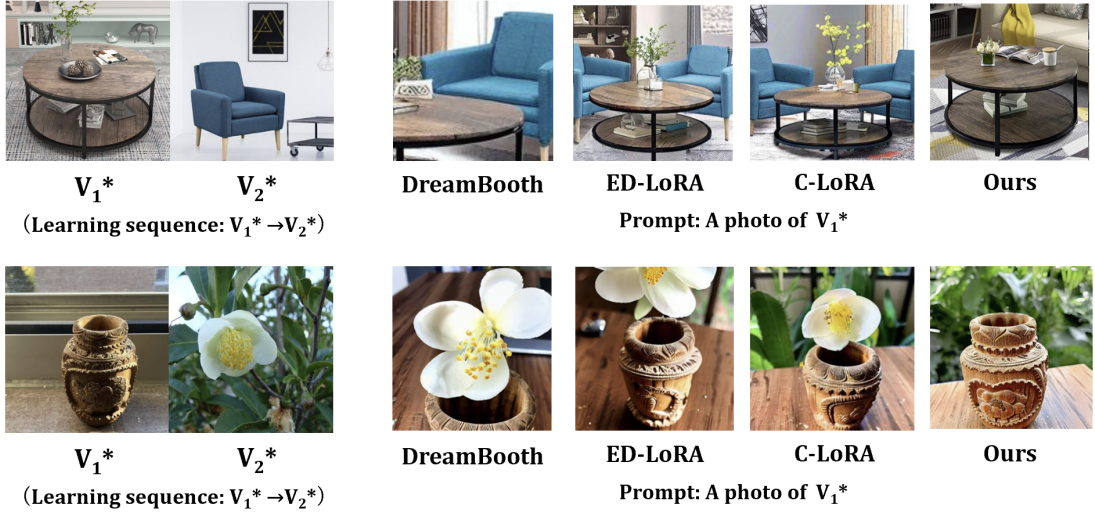


Figure 4.3: Qualitative results of experiments. All compared methods were affected by concept bleed-through in the learning sequence when dealing with both the nearby concepts (e.g., table and chair) and the related concepts (e.g., vase and flower). In contrast, our proposed method ensured accurate representation of the concepts in the image.

used stable diffusion v1.5 [19] as the backbone. In addition, we utilized the prompt “a photo of a V^* ,” where V^* is a learnable custom token. For LoRA, we searched for the rank using a simple exponential sweep and found that a rank of 16 learned all concepts sufficiently. During the training process, the learning rate was 2×10^{-6} , and the number of training steps performed on each image was set to 100.

In the optimization of the perturbation, we fixed the l_∞ -norm as the constraint to obtain all adversarial images. During the optimization process, we experimentally set the sampling step to 40, the total perturbation budget to $8/255$, and the batch size to 4.

Datasets: To evaluate the effectiveness of the proposed method, we used the official DreamBooth dataset [7], which comprises 30 distinct concept classes, each containing 5–6 images. Additionally, we also adopted the CustomConcept101 dataset [30], which comprises 101 distinct concept classes, each containing 3–15 images. These dataset are commonly used to train personalized diffusion models. Also, to optimize the adversarial perturbation, we employed target images sampled from ImageNet dataset [15]. **Compared methods:** The proposed method was compared with the following personalization methods for generative diffusion models: 1) Textual inversion [10], DreamBooth [7], InstantBooth [29]: baseline personalization methods for single concept; 2) Custom Diffusion [30], ED-LoRA [31]: mechanisms for training models that can

Table 4.1: Comparison of MMD and CLIP scores across different methods.

Method	$A_{mmd}(\downarrow)$	$F_{mmd}(\downarrow)$	CLIP Score($\uparrow$)
Textual inversion	2.98	0.00	0.71
DreamBooth	9.01	8.25	0.75
InstantBooth	3.34	6.91	0.74
Custom diffusion	5.57	3.96	0.78
ED-LoRA	7.72	7.08	0.82
C-LoRA	2.95	0.23	0.80
STAMINA	2.34	0.27	0.83
CIDM	2.11	0.22	0.85
C-LoRA + Ours	1.99	0.32	0.84
STAMINA + Ours	1.92	0.26	0.85
CIDM + Ours	1.98	0.22	0.87

generate multiple concepts simultaneously; 3) C-LoRA [23], STAMINA [2], CIDM [24]: SOTA of continual learning methods for personalized diffusion models.

Evaluation metrics: To assess the effectiveness of the proposed method, we employed evaluation metrics that focus on the quality of the generated images and the effectiveness of the perturbations in terms of preventing concept bleed-through. Specifically, we used Maximum mean discrepancy (MMD) score [71] and CLIP score [18] to evaluate the performance of the methods. Here, we defined A_{mmd} as the average MMD score ($\times 10^3$) across all concepts after completing the entire training sequence, and F_{mmd} measured the average increase of MMD score for the previously learned concepts after learning subsequent concepts, quantifying the degree of forgetting.

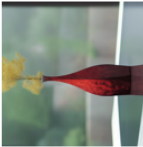
Training Step										
	0	1	2	3	4	5	6	7	8	9
	vase	cat_1	houseplant	cat_2	bag	teapot_1	teapot_2	can	teddybear	clock
Generation (Concept 0)	STAMINA									
										
Ours										
										

Figure 4.4: Examples of the additional experiment with different numbers of concepts in the training images. We trained the model with ten concepts in the learning sequence. We selected STAMINA [2] as the baseline method, and after each new concept was learned, the model was used to generate images of the most initially learned concept (*vase*).

Table 4.2: Performance of our method on different target images for the perturbation optimization of *chair*. “-” means we do not adopt any perturbation to the training images.

Target Image Category	Similarity with V1↓	Similarity with V2↓	A_{mmd} ↓	F_{mmd} ↓	CLIP ↑
-	-	-	3.12	0.51	0.76
vase	0.814	0.701	2.82	0.41	0.81
dog	0.187	0.325	1.92	0.25	0.85
airplane	0.523	0.544	1.96	0.28	0.84
bread	0.236	0.297	1.91	0.26	0.86
microwave	0.692	0.610	1.94	0.30	0.84

4.3.2 Experimental Results

Table 4.1 shows the quantitative results for each evaluated method. Note that all results are averaged over 100 different multiconcept learning patterns, where each pattern contains two sequentially learned personalization tasks. As can be seen, the proposed method achieved the highest CLIP score and the lowest A_{mmd} , demonstrating improved concept preservation and better alignment with textual prompts. In addition, in terms of the F_{mmd} results, the proposed method also performed well relative to preventing forgetting (note that textual inversion [10] has no forgetting because its backbone is completely frozen). These results suggest that adversarial perturbations can play a crucial role in enhancing the stability and reliability of diffusion-based continual learning models. It is also worth mentioning that our method exhibited significant results when applied to C-LoRA, STAMINA, and CIDM, proving that the proposed method is widely compatible with existing personalization frameworks. Figure 4.3 shows a qualitative comparison of the images generated using the different methods. As can be seen, the proposed method produced images that exhibit clearer concept boundaries and improved semantic alignment, particularly in challenging generation tasks.

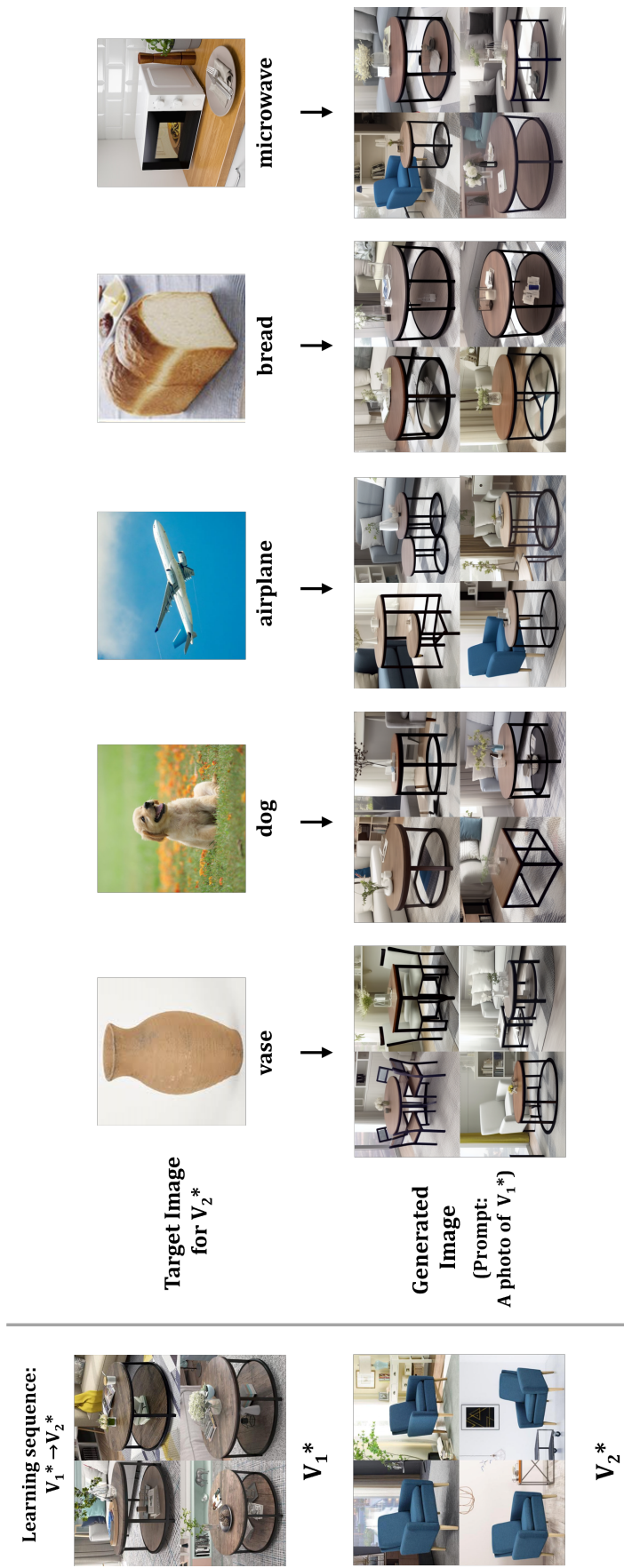


Figure 4.5: Examples of generated images using different target image categories for perturbation. Semantically dissimilar pairs lead to better concept separation.

Table 4.3: Comparison between selective forgetting methods and ours method.

Method	Chair CLIP Score $\uparrow$	Table CLIP Score $\uparrow$
C-LoRA	0.782	0.741
C-LoRA (w/SA)	0.851	0.094
C-LoRA (w/Ours)	0.847	0.822

4.3.3 Additional Experiments

In addition to the main experiments, we designed a series of additional experiments to further validate the effectiveness of the proposed method. Here, if not specifically stated, the experimental settings followed the main experiment. If not mentioned, we only applied our method to [24] in all additional experiments.

In the additional experiments, we manipulated the key variables as follows.

Number of Concepts in Training Images. A common challenge in continual learning for personalized image generation is the increasing risk of concept bleed-through as more concepts are introduced. To evaluate this, we followed the setup commonly seen in DreamBooth studies [7], where up to ten concepts are trained sequentially (in this experiment, the concept sequence was *vase*, *cat_1*, *houseplant*, *cat_2*, *bag*, *teapot_1*, *teapot_2*, *can*, *teddybear*, *clock*). The results are shown in Fig. 4.4. As can be seen, without appropriate mitigation strategies, concept interference becomes more pronounced as the number of concepts increases, leading to a degradation in the image generation quality. By integrating our adversarial perturbation strategy, we observed a significant improvement in terms of concept separation. Specifically, the images generated after learning all 10 concepts exhibited reduced semantic drift compared with the baseline methods without adversarial perturbations.

Categories of Target Images. The selection of the target image categories for adversarial perturbation optimization plays a crucial role in determining the effectiveness of the proposed method. Thus, we investigated different target image categories sampled from ImageNet, including images that do not overlap conceptually with the training data in the entire learning sequence and images that are semantically related to the training data. In this additional experiment, we employed learning sequence (*table* $\rightarrow$ *chair*) that was utilized in the main experiment. Specifically, we fixed the learning sequence to *table* $\rightarrow$ *chair*, and only inject perturbation to the training images for *chair*. Figure 4.5 shows the generation outcomes when training images for *chair*

are paired with target images from five different categories: *vase*, *dog*, *airplane*, *bread*, and *microwave* in the perturbation optimization. We also measured the cosine similarity between the generated image and both the *table* (V1) and *chair* (V2) embeddings.

Qualitative results (Fig. 4.5) indicate that pairing *chair* with highly dissimilar targets (e.g., *dog*, *bread*) yields clearer and more separable visual representations. Conversely, semantically closer pairs (e.g., *vase*, *microwave*) show subtle feature blending. These results align with the quantitative findings in Table 4.2. These results further validate the principle that semantic dissimilarity between training and target images promotes stronger adversarial-induced latent separation.

4.3.4 Comparison with Selective Forgetting Methods

In this section, we discuss works related to concept erasure and selective forgetting, which are closely connected to the topic of concept bleed-through. It is worth noting that while these methods share a similar objective (mitigating interference between concepts), their application scenarios and underlying philosophies differ significantly from our proposed method. Selective forgetting methods are typically designed to intentionally erase previously learned concepts before introducing a new one. This strategy is effective for avoiding interference but comes at the cost of losing the model’s ability to generate forgotten concepts. In contrast, our method does not require any forgetting steps and instead prevents concept bleed-through proactively during training. This allows the model to maintain the ability to generate all previously learned concepts, offering better flexibility and generalization in continual learning settings.

Despite these decisive differences, we still conducted additional comparative experiments against a representative forgetting-based method: Selective Amnesia (SA) [72]. Following the training sequence shown in Fig. 5.8 of our main manuscript (“*table* $\rightarrow$ *chair*”), we compared the quality of generation of *chair* after continual training. We use C-LoRA as the baseline method. In the case of Selective Amnesia, the model is explicitly fine-tuned to forget the concept of *table* before learning *chair*, whereas our method integrates adversarial perturbations to preserve separation without forgetting. Table 4.3 shows the CLIP similarity scores between generated images and their corresponding text prompts.

The results indicate that both methods achieve comparable performance on the new concept (*chair*). However, our method is able to retain the model’s ability to generate the previous concept

(*table*). This demonstrates that our method achieves a similar level of interference mitigation without sacrificing the knowledge of learned concepts.

4.4 Conclusion

This study has proposed an adversarial perturbation-based training strategy to mitigate the concept bleed-through issue in personalized generative models. By injecting a carefully designed adversarial perturbation into the training images, the proposed method shifts the semantic representations in the latent space, thereby ensuring a clear separation of the new and previously learned concepts. The proposed method was evaluated experimentally, and the results demonstrated the effectiveness of the method through contrastive learning and latent space alignment strategies, which help preserve the integrity of the learned concepts.

Chapter 5

Latent Representation-based Copyright Protection Methods for Artificially Created Works Using Adversarial Perturbation

5.1 Introduction

Recently, diffusion models [62] have been the focus of active research due to their ability to generate high-quality images. Text-to-image diffusion models, such as DALL-E 2 [63] and Stable Diffusion [19], have become the dominant approach in text-to-image generation. These models are used not only for generating images from arbitrary text but also for various tasks such as image editing [64, 65], super-resolution [66], inpainting [67, 68], and other related applications. Despite the extensive training of diffusion models, they may struggle to generate unique or personalized concepts that are absent from large-scale training datasets, such as specific faces or personalized styles. To address this, there has been a growing trend toward developing personalization techniques in text-to-image diffusion models, such as Textual Inversion [10] and DreamBooth [7]. These approaches allow for the creation of images related to a new, specific concept (object/style) using only a few reference examples. The learning process is highly efficient, as it requires modifying only a small portion of the weights in the diffusion model pipeline, enabling rapid concept acquisition and lightweight model updates.

Although research on personalized diffusion models is rapidly advancing, their security is a

critical issue that should not be overlooked. For example, similar to studies [73] on general text-to-image diffusion models, researchers have exposed vulnerability in personalized text-to-image diffusion models, making them susceptible to simple backdoor attacks [74]. By introducing mismatched text-image pairs during the pretraining process, attacker-defined triggers can be easily embedded into the model, leading to incorrect image generation for the given concept. These attack methods assume that the attacker can directly modify the denoising diffusion steps and upload pretrained models to third-party platforms, demonstrating strong attack performance in scenarios where users rely on pretrained models from such platforms. However, The risk of using pretrained models from third-party platforms is increasing because, unlike official models, unspecified users are often involved in the pretraining process.. In other words, while users can choose to avoid models from third-party platforms, the typical scenario of previous backdoor attacks become less practical in the context of personalized diffusion models.

A more effective type of attack in common scenarios using adversarial examples with imperceptible perturbations to disturb diffusion models' ability to generate proper image. This approach is model-agnostic, as it only modifies the input image, making it highly practical and generalizable. The attacker simply needs to upload a certain number of concept-specific adversarial examples to a public web platform. If a user collects training data from that platform to train a personalized text-to-image generation model and unknowingly includes the adversarial examples, then the attack can be executed. Recent methods for constructing adversarial examples targeting diffusion models have been proposed, demonstrating their effectiveness [3–6]. However, these methods only reduce the quality of the generated images and do not effectively prevent the diffusion model from learning the semantic information of the input images. While images generated using adversarial examples as input may exhibit low quality and significant noise, the underlying semantic information remains largely intact. This creates the potential for restoring noisy generated images to high-quality ones through image restoration or denoising techniques. Therefore, beyond simply degrading image quality, manipulating the semantic content of the generated images is crucial.

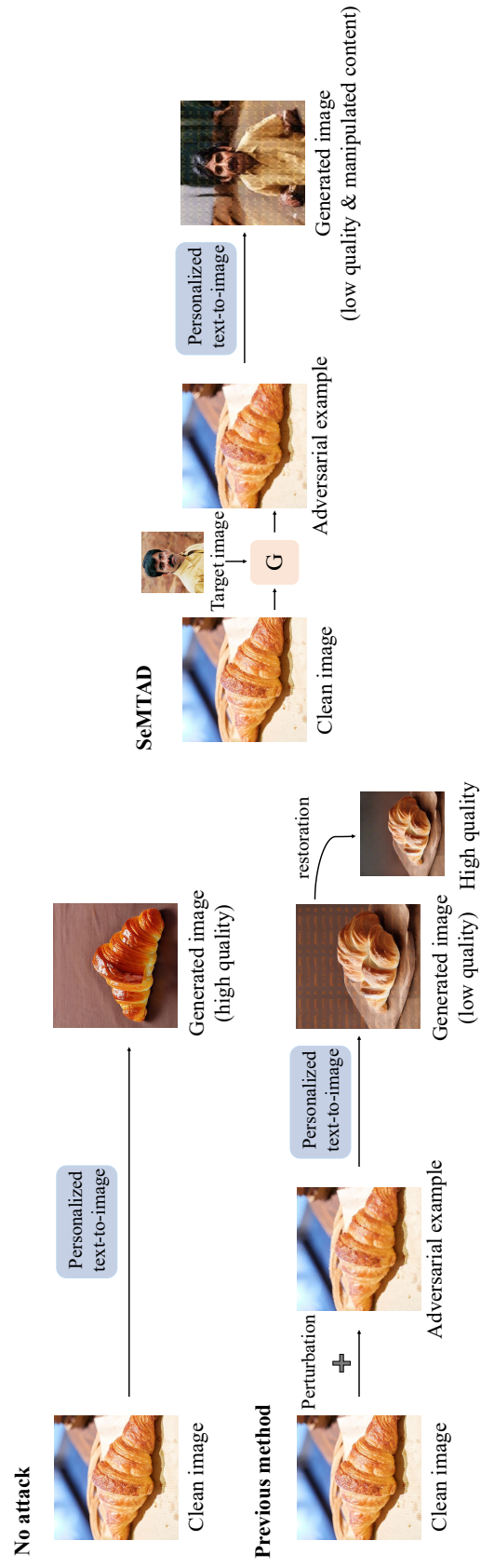


Figure 5.1: Examples of using the personalization method of the text-to-image diffusion model to generate images. Previous adversarial attack methods [3–6] introduce imperceptible perturbations to input images, causing personalized diffusion models trained on these images to generate outputs with significant noise or visible watermark. However, the underlying semantic information is still preserved, allowing simple image processing techniques to restore high-quality images. In contrast, the objective of our method, SeMTAD, is degrade the quality of the generated images but also to manipulate them such that the semantic information differs significantly from that of the input images. (In the figure, G represents the perturbation generation)

In this study, we constructed a semantic manipulation-based targeted adversarial attack for personalized text-to-image diffusion models (SeMTAD). We designed a perturbation generation algorithm that optimized adversarial attacks, accompanied by a novel loss function. This loss function ensures that adversarial examples are encoded as incorrect semantic representations, disrupting the diffusion model’s encoder from accurately capturing the semantic information within an image. To achieve precise manipulation of semantic content, we employ a semantic segmentation technique [75] to selectively target specific areas in the image where content modification is intended. Moreover, our adversarial examples effectively push the distribution of the generated model’s outputs far away from the distribution of real data, thereby impairing the model’s ability to generate high-quality and realistic images. As a result, the generated images exhibit reduced quality, and their semantic information can be manipulated. As shown in Fig. 5.1, compared to previous methods, our method provides a more powerful way to attack personalized text-to-image diffusion models without concerns about the restoration of the degraded images. The technical contributions of this study are as follows.

- Recognizing the limitations of traditional backdoor attacks for personalized diffusion models, we focus on adversarial attacks and introduce a novel approach that manipulates semantic information to construct adversarial examples for text-to-image diffusion models.
- We propose a perturbation generation algorithm that optimizes adversarial attacks, resulting in a significant semantic mismatch between the generated and input images.
- To enhance the accuracy of semantic manipulation, we utilize semantic segmentation to pinpoint and modify specific areas of the image content.
- Extensive experiments strongly demonstrate that SeMTAD effectively prevents personalized diffusion models from learning the semantic information of input images, with strong generalization across various mainstream personalization techniques.

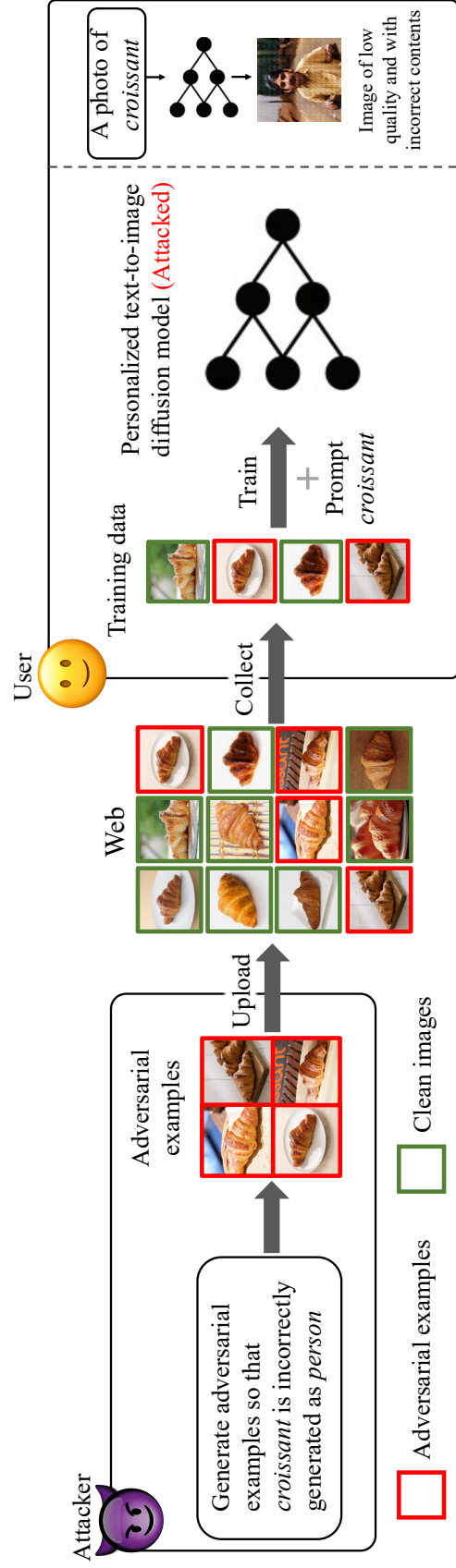


Figure 5.2: Our attack scenario. Users who collect training images from the Web may unknowingly include adversarial examples uploaded by the attacker. When this occurs, the personalized model will generate low-quality images with incorrect content, as specified by the attacker.

5.2 Preliminary

This section introduces diffusion models [62] and their working principles to clarify the problem and provide insights into the SeMTAD. In addition, we identified the threat model in our study.

5.2.1 Threat Model

Attack scenario. Training a text-to-image model from scratch is computationally expensive, so users often prefer personalization techniques that allow them to fine-tune a pre-trained model with just a few images. Targeting this, an attacker generates adversarial examples and uploads them to the Web. When users collect images for training from the Web, they risk unknowingly including these tampered adversarial examples. If the training data contain such examples, then the personalized model will produce low-quality images when users attempt to generate specific concepts. Moreover, the semantic content in the generated images will be altered to reflect the incorrect information embedded by the attacker. An overview of our attack scenario is shown in Fig. 5.2.

Attacker’s capability. In this study, we assume that the attacker can inject adversarial examples into the personalized model’s training dataset without having access to any other part of the model pipeline. Notably, unlike backdoor attacks on personalized text-to-image diffusion models, our approach does not require privileged access to the model’s training and deployment stages. Because users often collect images from the Web for the personalization process, our assumption aligns with real-world conditions, making this attack feasible for typical Internet users.

Attacker’s goal. By introducing adversarial examples into the training data of a personalized text-to-image diffusion model, the attacker seeks to cause the model to misbehave, i.e., generating low-quality images with incorrect semantic content.

5.2.2 Diffusion Process

Diffusion models employ a stochastic differential process known as the diffusion process. Starting with a sample $x_0 \sim q(\cdot)$, the diffusion process incrementally adds noise to generate samples $x_1, \dots, x_T$ for T steps, where $q(\cdot)$ represents the distribution of real data, $x_{t+1} = x_t + \varepsilon_t$

(t is sampled uniformly over the T time steps), and ε_t is sampled from the Gaussian distribution. As a result, the sample x_T starts to follow a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ when $T \rightarrow \infty$. Note that $\mathbf{I}$ is the identity matrix. If we can reverse this process and sample x_t from x_{t+1} , new samples from $q(\cdot)$ can be generated. This is achieved by simply starting from $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively denoising the samples for the T steps to produce a new image.

In other words, a denoising model ε_θ that predicts noise ε_t added to x_t at each time step t given x_{t+1} is required, where θ denotes the model parameter. Consequently, the denoising model ε_θ is trained to minimize the following loss function:

$$\mathcal{L}_{DM}(x, \theta) = \mathbb{E}_{t, \varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|\varepsilon - \varepsilon_\theta(x_t, t)\|_2^2 \right], \quad (5.1)$$

Eq. (5.1) is the core function of diffusion models; accordingly, our goal in adversarial attacks is to disrupt this process, thereby reducing the generated image quality.

5.2.3 Prompt-guided Sampling in Diffusion Models

By default, diffusion models generate random samples from the distribution of the training data $x_0 \sim q(\cdot)$. However, guiding the sampling process using natural language is feasible. For example, in latent diffusion models [19], which have demonstrated strong performance, the approach involves combining the latent representation of an input image, obtained through an encoder $\mathcal{E}$, with the embedding of user-defined text prompts. This allows the diffusion model to link the semantic information of the input image with specific text prompts, imparting control over the generated image. Therefore, perturbation becomes key in enabling adversarial attacks that target the semantic content of the generated image.

5.3 Methodology

In this section, we present the perturbation generation algorithm of SeMTAD. First, we outline the implementation of adversarial attacks on diffusion models (5.3.1). Next, we explain how SeMTAD incorporates semantic information into the attack (5.3.2). Finally, we provide the formulation of the loss function used to optimize SeMTAD (5.3.3). An overview of SeMTAD is shown in Fig. 5.3.

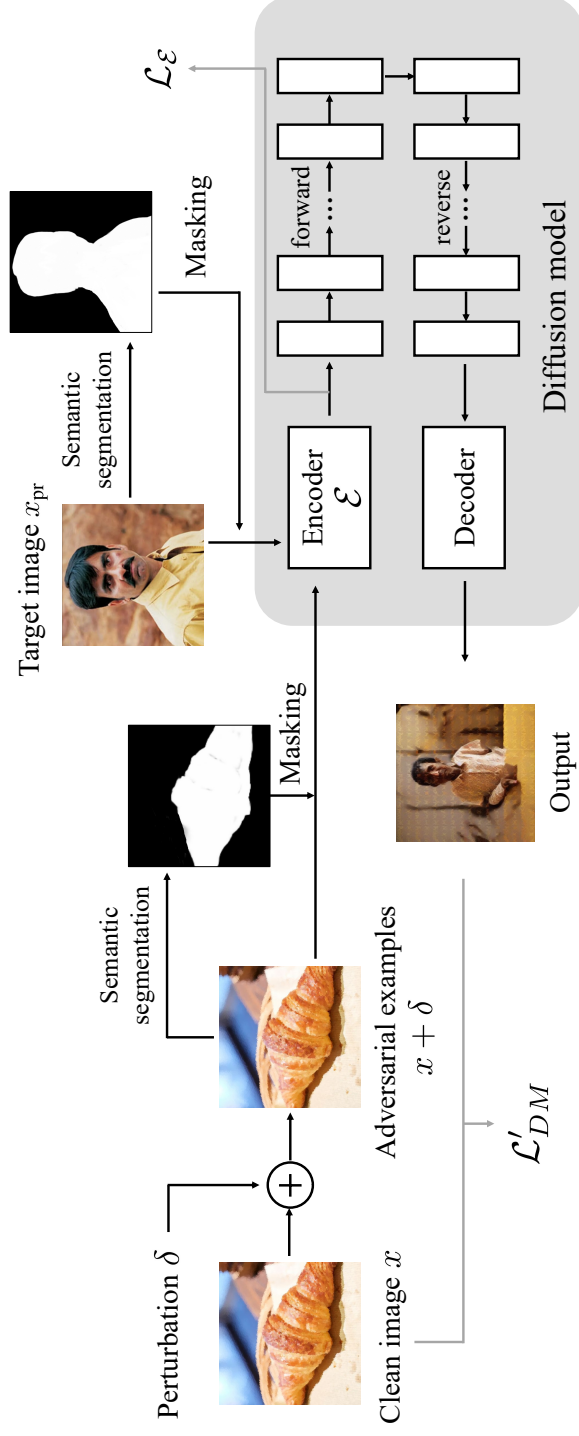


Figure 5.3: Our attack scenario. Users collecting images for training from the Web may unknowingly include adversarial examples uploaded by attackers. In this scenario, the personalized model will generate low-quality images with incorrect content, as dictated by the attacker.

5.3.1 Adversarial Attack for Diffusion Models

Several previous studies on adversarial attacks have shown promising results [76, 77]. However, most of these studies focused on image classification models, with relatively little research on adversarial attacks for generative models. Unlike image classification task, where the model learns to map images to categories, generative tasks involve learning to generate new images based on training data. Owing to the absence of explicit input–output pairs, such as image–category mappings, training a generative model by directly correcting its output based on the input is not feasible. Nevertheless, we aim to leverage the properties of generative tasks to define adversarial attacks specifically for them.

A generative model learns from real data $x \sim q(\cdot)$ and estimates a distribution $p_\theta(x)$ where it can sample generated data. To address the aforementioned lack of explicit data pairs, an intuitive idea to train a generative model is to maximize $p_\theta(x)$ for real data $x \sim q(\cdot)$ [17, 78].

However, $p_\theta(x)$ is difficult to directly optimize without applying a transformation. To overcome this, researchers have introduced a variational bound, represented by the following inequality:

$$\log p_\theta(x) \leq -\log \frac{p_\theta(x, z)}{q(z|x)}. \quad (5.2)$$

As state-of-the-art generative models, diffusion models follow this idea while constructing a series of latent variables $x_{1:T}$ by Markov chain $q(x_{1:T}|x_0)$. The reverse Markov chain $p_\theta(x_{0:T})$ is then used to revert the latent variables to the data x_0 , and $p_\theta(x)$ is optimized with the following variational bound of p_θ :

$$\log p_\theta(x) \leq -\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} := \mathcal{L}_{DM}. \quad (5.3)$$

This formula is equivalent to the loss function used to train the diffusion models.

We now consider adversarial attacks on diffusion models. As mentioned above, a common approach for training generative models is to maximize $p_\theta(x)$. Therefore, a similar idea is that if we provide an image $x' = x + \delta$ with a small perturbation δ such that $p_\theta(x')$ is minimized as much as possible, the generated images sampled from $p_\theta(x')$ will deviate significantly from real data. In other words, the generative model trained on x' will produce unrealistic and low-

quality images. On the basis of this idea, we define adversarial examples for a diffusion model parameterized by θ .

Definition: Given a diffusion model parameterized by θ and the distribution of real data $q(\cdot)$, an adversarial example x' is formulated by $x' = x + \delta$. δ is given as follows:

$$\delta := \arg \min_{\|\delta\| \leq \epsilon} p_\theta(x + \delta), \quad (5.4)$$

where $x \sim q(\cdot)$, and constant ϵ means the perturbation budget and is usually small.

To calculate p_θ in diffusion models, we expand $p_\theta(x)$ over the latent variable $x_{1:T}$ as follows:

$$p_\theta(x) = \int p_\theta(x_{0:T}) dx_{1:T}. \quad (5.5)$$

By minimizing $p_\theta(x_{0:T})$ with different sampling processes of $x_{1:T}$, we are approximately minimizing $p_\theta(x)$. Let $u(x'_{1:T})$ denotes the distribution of $x'_{1:T}$, and Eq. (5.4) can be rewritten as follows:

$$\delta := \arg \min_{\|\delta\| \leq \epsilon} \mathbb{E}_{x'_{1:T} \sim u(x'_{1:T})} p_\theta(x'_{0:T}). \quad (5.6)$$

A notable property of diffusion models is that the posterior distribution $q(x'_{1:T}|x'_0)$ is Gaussian. The parameters of this distribution are fixed and entirely independent of x'_0 . Therefore, this distribution remains invariant during the optimization of adversarial examples, allowing us to rewrite Eq. (5.6) as follows:

$$\begin{aligned} & \min_{\delta} \mathbb{E}_{x'_{1:T} \sim u(x'_{1:T})} p_\theta(x'_{0:T}) \\ &= \max_{\delta} \mathbb{E}_{x'_{1:T} \sim u(x'_{1:T})} -\log p_\theta(x'_{0:T}) \\ &= \max_{\delta} \mathbb{E}_{x'_{1:T} \sim u(x'_{1:T})} -\log \frac{p_\theta(x'_{0:T})}{q(x'_{1:T}|x'_0)} \\ &= \max_{\delta} \mathbb{E}_{x'_{1:T} \sim u(x'_{1:T})} \mathcal{L}_{DM}(x', \theta). \end{aligned} \quad (5.7)$$

Eq. (5.7) includes the loss function $\mathcal{L}_{DM}(x', \theta)$ as defined in Eq. (5.1); however, the optimization variables in Eqs. (5.1) and (5.7) are different. Specifically, when training diffusion models, the optimization variable is the model parameter θ . In contrast, when generating adversarial examples, the optimization variable is the perturbation δ added to the images. Further, by using

Eq. (5.1), we can rewrite the loss function in the following form:

$$\mathcal{L}'_{DM} = \mathbb{E}_{t, \varepsilon \sim \mathcal{N}(0,1)} \mathbb{E}_{x'_t \sim u(x_t)} \left[\left\| \varepsilon - \varepsilon_\theta(x'_t, t) \right\|_2^2 \right]. \quad (5.8)$$

By maximizing this loss function, we can generate adversarial examples that cause diffusion models trained on them to produce unrealistic and low-quality images.

5.3.2 Adversarial Attack Manipulating Semantic Information

The adversarial attack definition for diffusion models in subsection 5.3.1 emphasizes the goal of distancing generated images from real images. However, personalized text-to-image diffusion models, which learn the semantic information of images and associate them with specific text prompts, remain difficult to significantly perturb. As mentioned in subsection 5.2.3, we shift our focus to the encoding layer widely used in text-to-image diffusion models.

Using an image encoder $\mathcal{E}$, text-to-image diffusion models first encode input images into latent vector representations, which are then combined with text prompt embeddings to generate new images [19]. To execute an adversarial attack, we disrupt this process by forcing the encoder to map the input images to incorrect representations. To more accurately manipulate the semantic information in the image, we employ a semantic segmentation method [75] to create masks for the corresponding semantic parts of both the input and target images, and then extract the image features after masking. Specifically, we solve the following optimization problem using the projected gradient descent [70]:

$$\begin{aligned} \delta &:= \arg \min_{\|\delta\| \leq \epsilon} \mathcal{L}_{\mathcal{E}}(x, \delta, x_{\text{pr}}) \\ &= \arg \min_{\|\delta\| \leq \epsilon} \left\| \mathcal{E}(\mathcal{S}(x_{\text{pr}})) - \mathcal{E}(\mathcal{S}(x + \delta)) \right\|_2^2, \end{aligned} \quad (5.9)$$

where $\mathcal{S}$ represents the semantic segmentation mask, and x_{pr} is the target image specified by the user, making the attack controllable. In principle, x_{pr} should be semantically distant from x . According to Eq. (5.9) by minimizing the loss $\mathcal{L}_{\mathcal{E}}$, the encoder is unable to capture the correct latent representation of the input image. As a result, the subsequent combination with text prompts will be disturbed, causing the generated image to fail in reflecting the correct semantic information.

5.3.3 Final Loss

We combine the losses $\mathcal{L}'_{DM}$ and $\mathcal{L}_{\mathcal{E}}$ to formulate the final loss function $\mathcal{L}$. Specifically, we merge the two loss functions to create the following loss:

$$\begin{aligned}\mathcal{L} &= \lambda \mathcal{L}_{\mathcal{E}} - \mathcal{L}'_{DM} \\ &= \lambda \left\| \mathcal{E}(S(x_{pr})) - \mathcal{E}(S(x + \delta)) \right\|_2^2 \\ &\quad - \mathbb{E}_{t, \varepsilon \sim \mathcal{N}(0,1)} \mathbb{E}_{x'_t \sim u(x'_t)} \left[\left\| \varepsilon - \varepsilon_{\theta}(x'_t, t) \right\|_2^2 \right],\end{aligned}\tag{5.10}$$

where λ denotes the weight of $\mathcal{L}_{\mathcal{E}}$. By minimizing the combined loss $\mathcal{L}$, we generate adversarial examples that cause diffusion models trained on these examples to produce unrealistic and low-quality images, with a significant discrepancy between the semantic information of the generated and input images.

5.4 Experiments

5.4.1 Experimental Settings

Attack settings: We used Stable Diffusion 1.5 [19] as the backbone for our experiments. We fixed l_{∞} -norm as the constraint for obtaining all adversarial examples. During the training, we set the sampling step, total perturbation budget, and batch size to 40, 16/255, and 4, respectively—consistent with previous studies [4,5]. For the semantic segmentation in SeMTAD, we adopted the Segment Anything Model 2 (SAM 2) [75], which is one of the most effective and widely used models today.

To evaluate the effectiveness of our adversarial attack on personalized diffusion models, we targeted DreamBooth [7] and Textual Inversion [10], two of the most popular and powerful personalized text-to-image diffusion models, as the attack targets in our experiments. Specifically, DreamBooth and Textual Inversion fine-tune a pretrained text-to-image diffusion model, allowing specific object features to be associated with a given prompt ($[V]$), thereby enabling the generation of entirely new images containing those specific features. Similar to previous experiments, we also used Stable Diffusion [19] as the pre-trained generator. During our experiments, we re-trained both the U-Net and the text encoder, setting the learning rate to 2×10^{-6} and performing 100 training steps for each image.

Comparison methods: We selected three state-of-the-art attack methods for diffusion models to compare with SeMTAD.

- **Glaze** [3], a method primarily used to prevent diffusion models from learning specific artists’ styles. It uses style transformation tools to disrupt image features, after which subtle perturbations are optimized and added to the original image to execute the attack.
- **AdvDM** [4], a method we mimicked to calculate the loss function $\mathcal{L}'_{DM}$ in subsection 5.3.2. The main distinction between this method and SeMTAD lies in whether the semantic information of the generated images is manipulated.
- **Anti-DreamBooth** [5], a method specialized for attacking DreamBooth. It works by attacking the learning process of DreamBooth and forcing it to perform worse in reconstructing clean images. Because it is designed specifically for attacking DreamBooth, for fairness, we only compared AdvDM [4] and Glaze [3] in the experiments with Textual Inversion.

In addition, for some of the related methods whose source code has not yet been made public [6], we plan to experiment with them as comparison methods once their source code become publicly available.

Datasets: Following previous studies [4, 6], we selected the following datasets with category annotations:

LSUN [8], a dataset that contains approximately one million labeled images across 20 object categories (e.g., cat, airplane, and person), ImageNet [15], a large-scale image ontology built upon the WordNet structure, and UECFoodPix [9], a dataset comprising approximately 10,000 labeled food images for each of 103 categories (e.g., rice, pizza, and croissant).

In our experiments, we randomly selected 1,000 images each from LSUN-cat, ImageNet-cat, and UECFoodPix-croissant, respectively. The images were divided into groups of five. We first fine-tuned the diffusion model using DreamBooth on the images in each group and then conditionally generated 50 images, resulting in a total of 10,000 generated images per dataset. Specifically, the training instance and prior prompts were respectively “a photo of [V] animal” and “a photo of animal” when using ImageNet-cat/LSUN-cat and “a photo of [V] food” and “a photo of food” for UECFoodPix-croissant. By default, all images were resized to 256×256 .

This generation process was performed for both clean images and adversarial examples. To create adversarial examples for ImageNet-cat/LSUN-cat and UECFoodPix-croissant, we randomly selected images from LSUN-airplane and LSUN-person, respectively, as the target images mentioned in subsection 5.3.2, increasing the semantic distance between the input and target images. In addition, to verify the generalization of SeMTAD, we replaced DreamBooth with Textual Inversion and repeated the same process.

Evaluation metrics: Following previous studies [4], we evaluated the quality of the generated images using two metrics: Fréchet Inception Distance (FID) [79, 80], Precision (Prec.) [81], and Attack Success Rate (ASR).

FID is a metric for evaluating the quality of generated images and was introduced as an improvement over the Inception Score [82, 83]. A higher FID indicates that the generated images are more distant from the real images, meaning that, in our attack scenario, a higher FID corresponds to a more successful attack. In particular, although the metric Prec. is more commonly used for evaluating classification methods, it can also be applied to analyze the effectiveness of generative models from a distributional perspective. Researchers introduced the classic concepts of precision and recall to evaluate generative models, addressing the limitation that FID and related density metrics do not adequately capture these aspects [81]. A low FID may indicate either high precision (realistic images) or high recall (diverse images). Precision measures the fraction of generated images that are realistic, while recall quantifies the fraction of the training data manifold covered by the generator. Both are computed as expectations of binary set membership across distributions, specifically by measuring the probability that an image drawn from one distribution is classified as belonging to the support of another. Specifically, we denote the distribution of the real images with $\mathbf{P}_r$ and the distribution of the generated images with $\mathbf{P}_g$. Precision is defined as the probability that a random image from $\mathbf{P}_g$ falls within the support of $\mathbf{P}_r$, while recall is the probability that a random image from $\mathbf{P}_r$ falls within the support of $\mathbf{P}_g$. Because SeMTAD is designed to attack few-shot personalized text-to-image models, recall, which measures the diversity similarity between the generated and real images, is less relevant in our context. Therefore, in the experiments validating the effectiveness of SeMTAD, we used only precision as the evaluation metric.

Table 5.1: Experimental results for DreamBooth [7].

Method	LSUN-cat [8]			ImageNet-cat [15]			UECFoodPix-croissant [9]		
	FID (↑)	Prec. (↓)	ASR (↑)	FID (↑)	Prec. (↓)	ASR (↑)	FID (↑)	Prec. (↓)	ASR (↑)
No Attack	23.78	0.579	0.000	29.13	0.567	0.000	34.94	0.530	0.000
Glaze [3]	145.26	0.174	0.819	143.65	0.190	0.723	122.31	0.273	0.791
AdvDM [4]	155.77	0.150	0.867	143.56	0.182	0.757	146.35	0.198	0.827
Anti-DreamBooth [5]	140.23	0.207	0.810	131.22	0.212	0.761	133.89	0.246	0.798
SeMTAD	189.12	0.073	0.962	192.56	0.104	0.943	198.72	0.059	0.991

Table 5.2: Experimental results for Textual Inversion [10].

Method	LSUN-cat [8]			ImageNet-cat [15]			UECFoodPix-croissant [9]		
	FID (↑)	Prec. (↓)	ASR (↑)	FID (↑)	Prec. (↓)	ASR (↑)	FID (↑)	Prec. (↓)	ASR (↑)
No Attack	22.12	0.534	0.000	21.91	0.539	0.000	32.90	0.513	0.000
AdvDM [4]	176.46	0.128	0.891	153.57	0.164	0.814	178.12	0.119	0.894
Glaze [3]	159.73	0.175	0.874	143.63	0.186	0.791	144.84	0.202	0.823
SeMTAD	210.34	0.032	0.963	203.40	0.087	0.934	230.89	0.011	0.996

Table 5.3: Results of ablation studies on different category pairs. We set the number of adversarial examples in the training data to 5. The attack target is DreamBooth [7].

Category	FID ($\uparrow$)	Prec. ($\downarrow$)	ASR ($\uparrow$)
cat (no attack)	23.78	0.579	0.000
cat $\rightarrow$ airplane	189.12	0.073	0.962
cat $\rightarrow$ dog	198.23	0.069	0.984
car (no attack)	21.56	0.623	0.000
car $\rightarrow$ airplane	168.24	0.081	0.926
car $\rightarrow$ dog	171.98	0.074	0.955

Moreover, because SeMTAD manipulates the semantic content of the generated images, we also incorporate the ASR [74]. To calculate ASR, we first generated N_G images using the attacked model, then classified these images using CLIP [18]. The ASR is computed as $ASR = (N_G - N_P)/N_G$, where N_P represents the number of images classified as resembling the original image. A higher ASR indicates a more effective adversarial attack.

5.4.2 Experimental Results

Herein, we present the quantitative evaluation results in Tables 5.1 and 5.2. As shown in the tables, SeMTAD outperforms the comparison methods, achieving significantly higher FID and lower Prec. on both datasets compared to other personalized text-to-image generation methods. In addition, a comparison between SeMTAD with AdvDM [4] reveals that incorporating semantic information enhances attack performance. Figure 5.4 and 5.5 shows the qualitative evaluation results, where images generated by the diffusion model trained with our adversarial examples exhibit not only of poor quality (blurred and containing significant noise) but also altered semantic content, like the adversarial examples themselves. In contrast, other methods fail to achieve this semantic shift, indicating that SeMTAD demonstrates superior attack performance.

5.4.3 Effectiveness Evaluation

In addition to the primary experiments, we conducted a series of supplementary experiments to further validate the effectiveness of SeMTAD. Unless otherwise specified, the experimental settings remain consistent with those used in the main experiments.

Table 5.4: Results of the ablation studies on different category pairs. We set the number of adversarial examples included in the training data to 5. The target of the attack is Textual Inversion [10].

Category	FID ($\uparrow$)	Prec. ($\downarrow$)	ASR ($\uparrow$)
cat (no attack)	25.69	0.531	0.000
cat→airplane	198.77	0.061	0.975
cat→dog	210.54	0.051	0.991
car (no attack)	23.61	0.584	0.000
car→airplane	174.39	0.078	0.933
car→dog	183.76	0.074	0.964

5.4.3.1 Ablation Studies

In the ablation studies, we manipulated the key variables as follows:

The number of adversarial examples included in the training data for diffusion models.

Because, in our attack scenario, the number of adversarial examples ultimately incorporated into the training data for diffusion models is uncertain. To address this, we conducted ablation studies by varying the number of adversarial examples in the training data to validate the effectiveness of our attack under different conditions. Specifically, following previous studies [7, 74], we experimented on LSUN-cat, using 5 training images and varying the number of adversarial examples from 0 to 5. The target images were drawn from the LSUN-airplane. The results are shown in Fig. 5.6. Evidently, when two or more training images are adversarial examples, a significant improvement in attack performance is observed. Figure 5.7 provides an example from this experiment, demonstrating that not all training images need to be adversarial. In other words, even a small number of adversarial examples mixed into the training data can be highly effective.

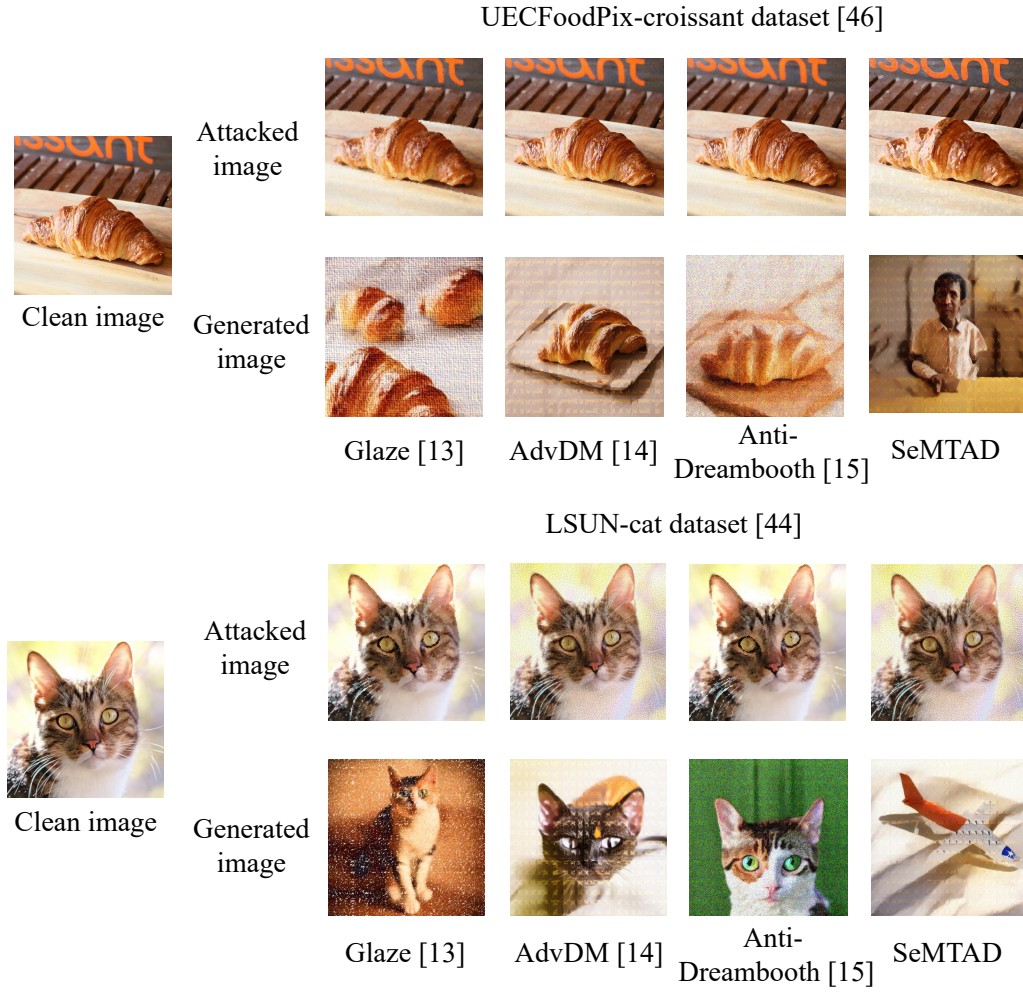


Figure 5.4: Attack results against DreamBooth [7]. We compared the images generated by SeMTAD and the benchmark methods on two datasets, LSUN-cat [8] and UECFoodPix-croissant [9]. SeMTAD successfully altered the content of the images, transforming a cat into an airplane and a croissant into person, respectively.

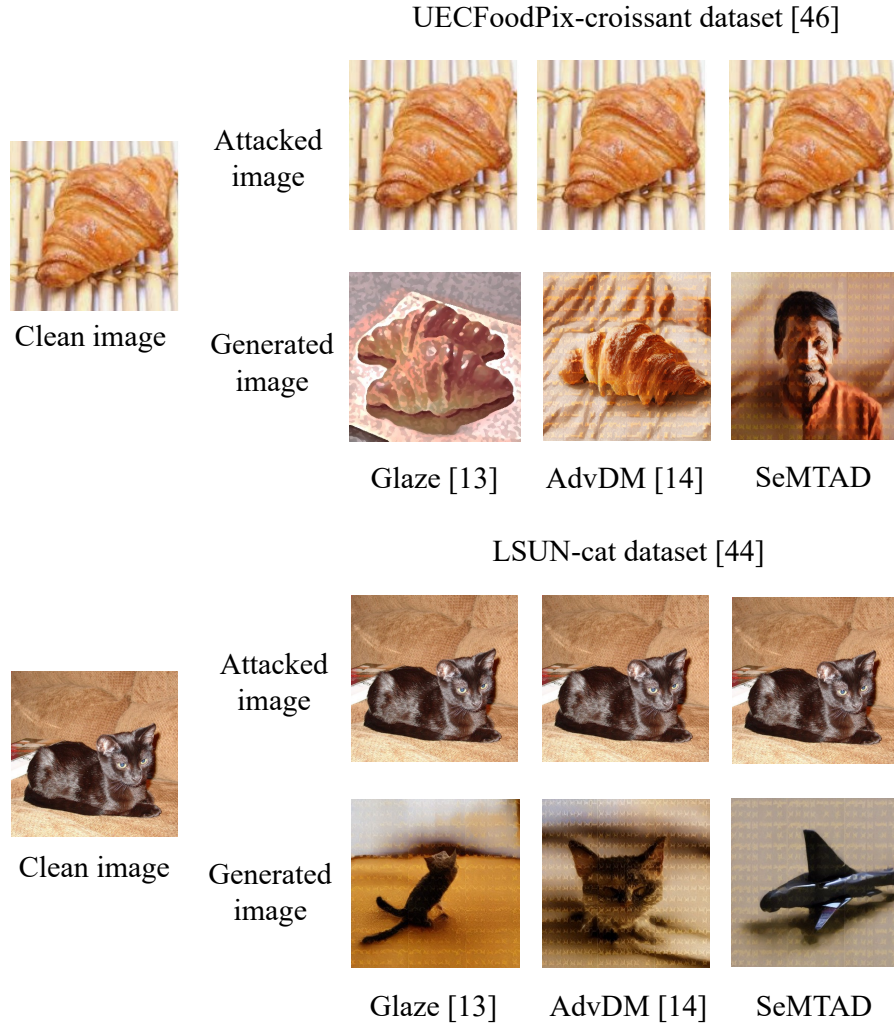


Figure 5.5: Attack results against Textual Inversion [10]. We compared the images generated by SeMTAD and the benchmark methods on two datasets, LSUN-cat [8] and UECFoodPix-croissant [9]. SeMTAD successfully altered the content of the images, transforming a cat into an airplane and a croissant into person, respectively.

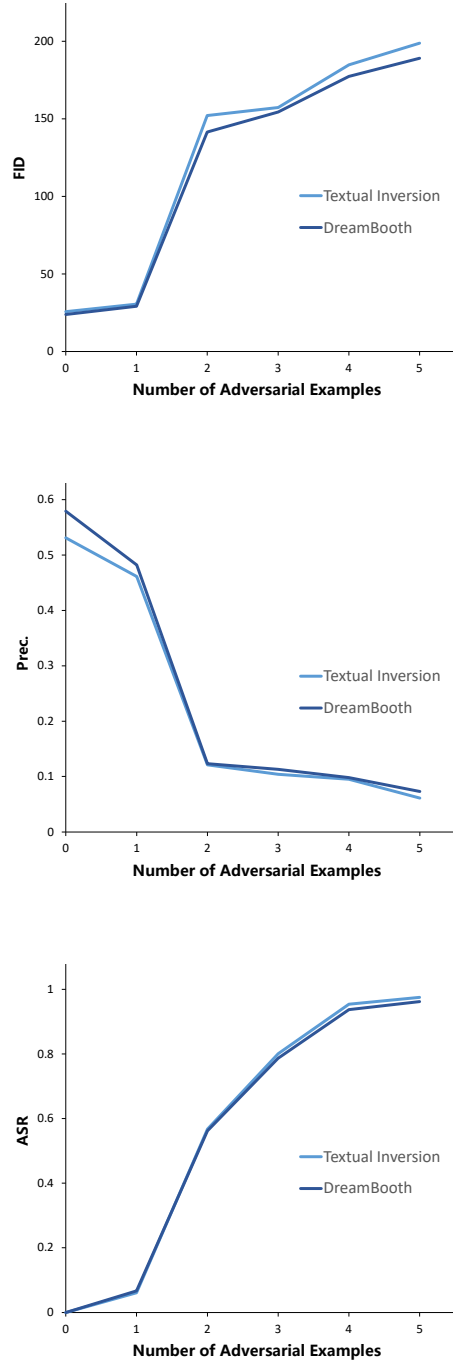


Figure 5.6: Results of ablation studies on LUSN-cat [8] with respect to changing the number of adversarial examples included in training data.

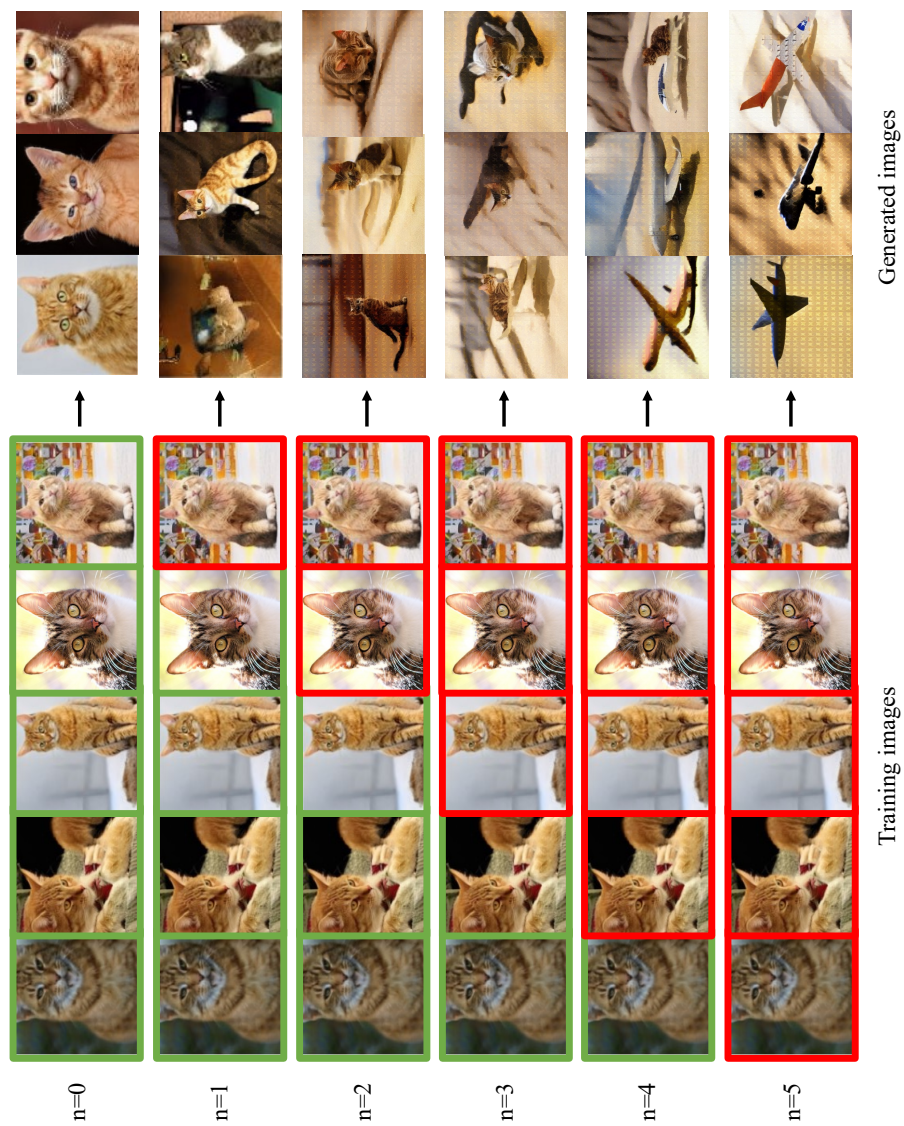


Figure 5.7: Example of the ablation study, illustrating the impact of varying the number of adversarial examples in the training data. Green boxes represent clean images, while red boxes indicate adversarial examples included in the training data. The attack target is DreamBooth [7].



Figure 5.8: Examples of ablation studies on category pair cat→dog and car→airplane. All training images are adversarial examples. The attack target is DreamBooth [7].

Categories of Target Images. Because the choice of target images significantly influences the performance of SeMTAD, we selected four pairs of categories of input and target image categories for experiments on LSUN: cat→airplane, cat→dog, car→airplane, and car→dog (where the input image category is on the left side of the arrow, and the attacker-determined target category is on the right). The number of adversarial examples in the training data was set to 5. The results, shown in Table 5.3 and 5.4, demonstrate that the ASRs across different category pairs are consistently close to 1, highlighting the exceptional performance of SeMTAD. Figure 5.8 provides examples from the experiment, showing that SeMTAD achieves strong attack performance across various category pairs.

Perturbation Budget. Because the perturbation budget is a crucial parameter for generating adversarial examples, we investigated its impact on the quality of the adversarial example produced by SeMTAD. In this experiment, the perturbation budget ϵ was varied from $2/255$ to $32/255$. Tables 5.5 and 5.6 show the results. We observe that even with a small perturbation budget ($8/255$), SeMTAD is able to significantly degrade the quality of images generated by both DreamBooth and Textual Inversion, without excessively compromising the quality of the original images.

Table 5.5: Effects of our attack under different perturbation budgets. The target of the attack is DreamBooth [7].

Perturbation budget	LSUN-cat [8]			ImageNet-cat [15]			UECFoodPix -croissant [9]		
	FID ($\uparrow$)	Prec. ($\downarrow$)	ASR ($\uparrow$)	FID ($\uparrow$)	Prec. ($\downarrow$)	ASR ($\uparrow$)	FID ($\uparrow$)	Prec. ($\downarrow$)	ASR ($\uparrow$)
$\epsilon = 0$ (no attack)	23.78	0.579	0.000	29.13	0.567	0.000	34.94	0.530	0.000
$\epsilon = 2/255$	25.59	0.494	0.003	31.40	0.487	0.001	35.28	0.418	0.002
$\epsilon = 4/255$	38.12	0.278	0.146	39.63	0.263	0.133	40.39	0.264	0.155
$\epsilon = 8/255$	144.46	0.107	0.827	151.53	0.105	0.845	156.31	0.096	0.842
$\epsilon = 16/255$	189.12	0.073	0.962	192.56	0.104	0.943	198.72	0.059	0.971
$\epsilon = 32/255$	205.68	0.019	0.995	208.79	0.017	0.997	212.04	0.016	0.997

Table 5.6: Effects of our attack under different perturbation budgets. The target of the attack is Textual Inversion [10].

Perturbation budget	LSUN-cat [8]			ImageNet-cat [15]			UECFoodPix -croissant [9]		
	FID ($\uparrow$)	Prec. ($\downarrow$)	ASR ($\uparrow$)	FID ($\uparrow$)	Prec. ($\downarrow$)	ASR ($\uparrow$)	FID ($\uparrow$)	Prec. ($\downarrow$)	ASR ($\uparrow$)
$\epsilon = 0$ (no attack)	25.69	0.531	0.000	21.91	0.539	0.000	37.04	0.517	0.000
$\epsilon = 2/255$	27.33	0.502	0.004	25.81	0.511	0.000	39.28	0.482	0.003
$\epsilon = 4/255$	40.03	0.262	0.194	39.66	0.272	0.117	43.87	0.251	0.143
$\epsilon = 8/255$	152.50	0.096	0.878	154.96	0.094	0.824	169.43	0.081	0.889
$\epsilon = 16/255$	196.10	0.042	0.969	203.40	0.087	0.934	203.35	0.037	0.984
$\epsilon = 32/255$	226.49	0.012	0.998	229.98	0.009	0.997	234.56	0.007	0.999

Table 5.7: Results of ablation studies on ImageNet-cat [15] with and without incorporating semantic segmentation (SAM2) into SeMTAD. The attack target is DreamBooth [7].

Method	FID ($\uparrow$)	Prec. ($\downarrow$)	ASR ($\uparrow$)
SeMTAD (w/ SAM2)	192.56	0.104	0.943
SeMTAD (w/o SAM2)	184.31	0.139	0.924

Utilization of Semantic Segmentation. To assess whether incorporating semantic segmentation improves the effectiveness of the attack, we conducted an ablation study by removing the semantic segmentation component from SeMTAD’s perturbation optimization algorithm and directly extracting the representation of the whole image. Table 5.7 shows the results for the ImageNet-cat dataset. The findings reveal that, after removing the semantic segmentation and masking processes, the attack performance deteriorates. This shows that semantic segmentation plays an indispensable role in enhancing the attack’s effectiveness.

5.4.3.2 Robustness Validation

We performed robustness validation of our adversarial attack. Following a previous study [4], we applied JPEG compression [11], TVM [12], and SR [13] as defense methods to our adversarial examples. The results are shown in Fig. 5.9. While preprocessing defenses can partially mitigate the effects of SeMTAD, they fail to fully restore the semantic information of the training images.

5.4.3.3 Validation of Image Fidelity

An adversarial attack disrupts the model’s proper operation; however, preserving the fidelity of the image after applying the perturbation is crucial. Although SeMTAD aims to minimize the perturbation, additional experimental results are needed to support this claim.

To ensure an objective and fair judgment, we randomly selected 1,000 images from both the LSUN-cat [8] and UECFoodPix-croissant [9] datasets and processed them using SeMTAD, as well as the comparison methods Glaze [3], AdvDM [4], and Anti-DreamBooth [5]. The perturbation budget ϵ for all methods was set to 16/255. Then, we calculated the average PSNR and SSIM [84] of the perturbed images relative to the original images. Table 5.8 shows the experimental results. As shown in the table, the impact of SeMTAD on image quality is comparable to that of the other methods. Not only do the qualitative results in the main text demonstrate minimal differences in image quality between SeMTAD and the other methods, but also this is also confirmed through quantitative analysis. Notably, while SeMTAD effectively manipulates

the semantic content of the generated images, previous methods fail to achieve this. Despite executing a more advanced attack, SeMTAD maintains the same level of image fidelity as other existing approaches. Therefore, the effectiveness of SeMTAD was validated.

5.5 Limitations

To validate the effectiveness of SeMTAD by considering the complexity of real-world application scenarios, we conducted an ablation study by changing the categories of input and target images. However, we randomly selected only a few category pairs, without accounting for the semantic distance between these categories. In reality, when attacking images in the category of *cat*, there should be a difference in the difficulty of forcing the generated images into *dog* and *airplane*, because the semantic distance between *cat* and *dog* is smaller than that between *cat* and *airplane*. This aspect warrants further exploration in future work. Moreover, for any given category of input images, there may exist specific target images that optimize the effectiveness of the attack with minimal modifications to the original images. In other words, this involves optimizing the selection of target images, a problem we plan to address in future research.

5.6 Conclusions

In this study, we developed an adversarial attack method based on semantic manipulation for personalized text-to-image diffusion models (SeMTAD). By forcing the encoder of diffusion models to map an input image to incorrect semantic representations, our attack creates a significant discrepancy between the semantic content of the generated and input images. Extensive experiments have demonstrated that SeMTAD effectively prevents diffusion models from learning the correct semantic information of the input image.

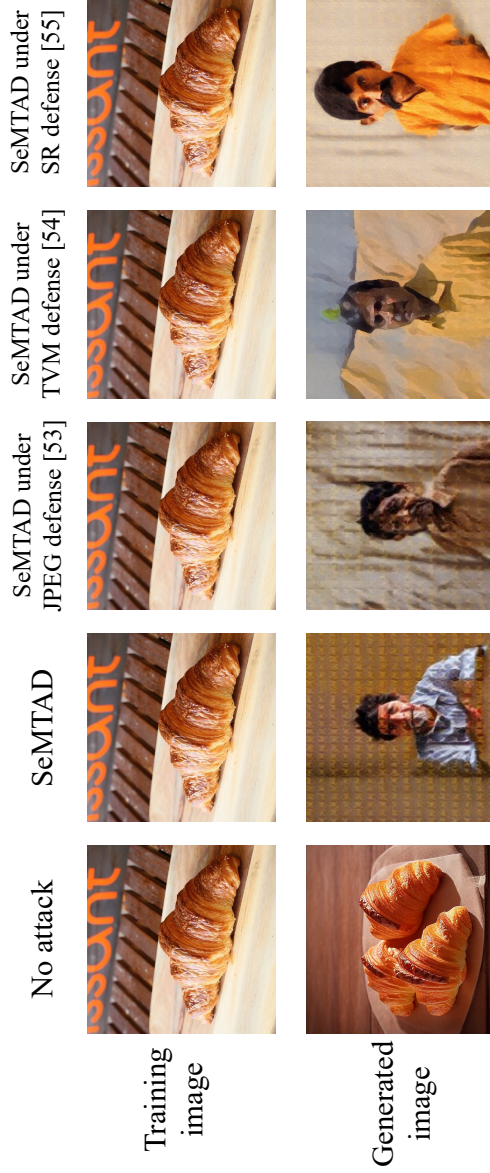


Figure 5.9: Results of the robustness validation on category pairs croissant $\rightarrow$ person. We applied JPEG compression [11], TVM [12], and SR [13] as defense methods on our adversarial examples. The attack target is DreamBooth [7].

Table 5.8: Experimental results of the validation of the image fidelity.

Method	ImageNet-cat [15]				UECFoodPix-croissant [9]			
	Adversarial example		Generated image		Adversarial example		Generated image	
	PSNR($\uparrow$)	SSIM($\uparrow$)	FID($\uparrow$)	ASR($\uparrow$)	PSNR($\uparrow$)	SSIM($\uparrow$)	FID($\uparrow$)	ASR($\uparrow$)
Glaze [3]	38.88	0.943	143.65	0.723	39.58	0.954	122.31	0.791
AdvDM [4]	38.12	0.920	143.56	0.757	41.69	0.964	146.35	0.827
Anti-DreamBooth [5]	40.48	0.951	131.22	0.761	41.53	0.961	133.89	0.798
SeMTAD	41.12	0.955	192.56	0.943	41.23	0.958	198.72	0.991

Chapter 6

Conclusions

As conclusions of this thesis, this chapter reviews the overview of the proposition and shows future directions.

6.1 Overview of the Proposition in This Thesis

This section provides an overview of the proposition presented in this thesis, building upon the background and previous chapters.

The proposition of this thesis is to establish latent representation learning as a unified framework for understanding, generating, and protecting artificially created works. To achieve this goal, the thesis investigates three fundamental technical challenges and proposes dedicated methods for each of them. The first challenge is to improve the semantic understanding of artificially created works, which often contain complex multi-label and hierarchical attribute structures. To address this, the thesis introduces latent representation-based recognition methods that incorporate hierarchical dependency modeling, multi-modal integration, and graph-based relational learning. These techniques allow models to extract structured semantics from stylized or non-photorealistic visual domains.

Building upon semantic understanding, the second challenge focuses on achieving robust generative controllability in personalized text-to-image diffusion models. Existing personalization methods tend to suffer from concept interference during continual learning, leading to degraded representation quality across concepts. To overcome this issue, the thesis proposes an adversarial perturbation-based continual personalization method that regulates latent semantic representations in the diffusion process. This approach strategically manipulates the latent space to ensure that newly learned concepts remain distinct from previously acquired ones, thereby preventing concept bleed-through while maintaining high generative fidelity.

Lastly, the third challenge is the protection of human-created works from unauthorized training by generative models. The thesis presents latent-space-oriented adversarial perturbation methods that alter sensitive semantic representations while preserving human-perceptual quality. These methods prevent diffusion models from learning stylistic or conceptual patterns embedded in artworks, thereby strengthening copyright protection. With the incorporation of the three lines of methods—semantic understanding, generative controllability, and copyright protection—the latent representation learning framework introduced in this thesis effectively addresses the major open problems in the field. Below, an overview of each chapter of this thesis is reviewed.

In Chapter 2, related works on semantic understanding, personalized diffusion models, and copyright protection are presented, and the problems addressed in this thesis are clarified. Chapter 3 presents latent representation-based semantic understanding methods for artificially created works, focusing on hierarchical attribute modeling and multi-modal semantic integration. Chapter 4 proposes an adversarial perturbation-based continual personalization method for diffusion models to mitigate concept bleed-through. Chapter 5 presents latent-space-oriented adversarial protection methods that prevent unauthorized training of generative models while maintaining visual fidelity.

The contributions of this thesis can be summarized as follows:

- The thesis introduces a unified latent representation learning framework that addresses semantic understanding, generative controllability, and content protection for artificially created works. The proposed methods improve the interpretability and effectiveness of latent semantic modeling.
- The thesis proposes an adversarial perturbation-based continual personalization method for diffusion models, which successfully mitigates concept bleed-through and stabilizes new concept learning in multi-concept settings.
- The thesis presents latent-space-oriented perturbation techniques for protecting artworks from unauthorized training, contributing practical solutions to copyright protection in the era of generative AI.

6.2 Future Directions

This section describes the future directions of this study. Future work can focus on refining and expanding the proposed latent representation learning methods. This includes enhancing semantic modeling techniques to capture more diverse and fine-grained attributes, especially in highly stylized or abstract artificially created works. Improving the robustness of hierarchical and relational modeling against domain shifts or varying artistic styles is also an important direction.

Similarly, continual personalization for diffusion models can be further improved by exploring more advanced latent-space regularization strategies. Developing mechanisms that explicitly separate shared and concept-specific semantic components may lead to even more stable multi-concept learning. Additionally, the integration of replay-based or contrastive techniques may enhance long-term concept retention. Since personalized generation is expected to play an increasingly important role in creative workflows, improving its stability and safety will remain a crucial research challenge.

Future research can also extend the proposed protection methods. While adversarial perturbation techniques demonstrated strong effectiveness, improving their robustness across diverse model architectures, preprocessing pipelines, and fine-tuning strategies is an important direction. Investigating protection strategies that adapt dynamically to different threat models or evolving generative architectures will be essential for long-term applicability. Furthermore, combining latent-space perturbation with watermarking or authentication techniques may lead to more comprehensive protection systems.

As latent representation learning continues to influence the fields of generation, understanding, and protection, it is important to consider the broader ethical and societal implications. Future work can explore guidelines for responsible personalization, limitations of automated creativity, and the balance between openness and protection in model development. Developing frameworks that ensure fairness, transparency, and accountability in the handling of artificially created works will be crucial for their safe and beneficial use across industries.

Acknowledgements

First, I wish to express my deepest gratitude to my supervisors, Prof. Miki Haseyama and Prof. Takahiro Ogawa. Throughout my five-year tenure at the Graduate School of Information Science and Technology, Hokkaido University, their invaluable guidance and unwavering support have played a crucial role in fostering my growth as a researcher. They are respected and charismatic mentors who have exerted a profound influence on my scientific career. I would also like to express my sincere appreciation to Prof. Yoshinori Dobashi for his insightful comments and constructive suggestions, which significantly refined my research.

I also extend my sincere appreciation to Associate Professor Ren Togo, Associate Professor Keisuke Maeda, Specially Appointed Assistant Professor Guang Li, and Assistant Professor Naoki Saito. I am truly grateful for their constant support and professional advice regarding both my research endeavours and academic life. In particular, I would like to pay high tribute to Associate Professor Keisuke Maeda. His patience and strong sense of responsibility have deeply shaped my character and my attitude towards life and research.

Finally, I would like to thank my family for their unsparing support. Their silent dedication and selfless assistance have been the bedrock of my life's journey. My thanks also go to all members of the Laboratory of Media Dynamics for their care and camaraderie, and to Hokkaido University for providing the excellent platform and resources that made this work possible.

References

- [1] T. D. Community and G. Branwen, “Danbooru2020: A large-scale crowdsourced and tagged anime illustration dataset,” <https://www.gwern.net/Danbooru2020>, 2021.
- [2] J. S. Smith, Y.-C. Hsu, Z. Kira, Y. Shen, and H. Jin, “Continual diffusion with stamina: Stack-and-mask incremental adapters,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1744–1754.
- [3] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao, “Glaze: Protecting artists from style mimicry by text-to-image models,” in *Proc. USENIX Security Symposium*, 2023.
- [4] C. Liang, X. Wu, Y. Hua, J. Zhang, Y. Xue, T. Song, Z. Xue, R. Ma, and H. Guan, “Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples,” *arXiv preprint arXiv:2302.04578*, 2023.
- [5] T. Van Le, H. Phung, T. H. Nguyen, Q. Dao, N. Tran, and A. Tran, “Anti-dreambooth: Protecting users from personalized text-to-image synthesis,” *arXiv preprint arXiv:2303.15433*, 2023.
- [6] P. Zhu, T. Takahashi, and H. Kataoka, “Watermark-embedded adversarial examples for copyright protection against diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24420–24430.
- [7] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, “Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation,” *arXiv preprint arxiv:2208.12242*, 2022.
- [8] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, “LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop,” *arXiv preprint arXiv:1506.03365*, 2015.
- [9] T. Ege and K. Yanai, “A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice,” in *Proc. ACM MM Workshop on Multimedia Assisted Dietary Management*, 2019.

- [10] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, “An image is worth one word: Personalizing text-to-image generation using textual inversion,” *arXiv preprint arXiv:2208.01618*, 2022.
- [11] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, S. Li, L. Chen, M. E. Kounavis, and D. H. Chau, “Shield: Fast, practical defense and vaccination for deep learning using jpeg compression,” in *Proc. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 196–204.
- [12] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, “Countering adversarial images using input transformations,” *arXiv preprint arXiv:1711.00117*, 2017.
- [13] A. Mustafa, S. H. Khan, M. Hayat, J. Shen, and L. Shao, “Image super-resolution as a defense against adversarial attacks,” *IEEE Transactions on Image Processing*, vol. 29, pp. 1711–1724, 2019.
- [14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [17] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, “Challenging common assumptions in the unsupervised learning of disentangled representations,” in *Proc. International Conference on Machine Learning*, 2019, pp. 4114–4124.
- [18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, 2021, pp. 8748–8763.
- [19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.
- [20] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

- [21] S. Banik, M. Lauri, and S. Frintrop, “Multi-label object attribute classification using a convolutional neural network,” *arXiv preprint arXiv:1811.04309*, 2018.
- [22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., “LoRA: Low-rank adaptation of large language models.,” *ICLR*, vol. 1, no. 2, pp. 3, 2022.
- [23] J. S. Smith, Y.-C. Hsu, L. Zhang, T. Hua, Z. Kira, Y. Shen, and H. Jin, “Continual diffusion: Continual customization of text-to-image diffusion with c-lora,” *arXiv preprint arXiv:2304.06027*, 2023.
- [24] J. Dong, W. Liang, H. Li, D. Zhang, M. Cao, H. Ding, S. H. Khan, and F. Shahbaz Khan, “How to continually adapt text-to-image diffusion models for flexible customization?,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 130057–130083, 2024.
- [25] T. Chen, Xu, Muxin, X. Hui, H. Wu, and L. Lin, “Learning semantic-specific graph representation for multi-label image recognition,” in *Proc. IEEE International Conference on Computer Vision*, 2019, pp. 522–531.
- [26] Z. Chen, X.-S. Wei, P. Wang, and Y. Guo, “Multi-label image recognition with graphconvolutional network,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5177–5186.
- [27] J. Ye, J. He, X. Peng, W. Wu, and Y. Qiao, “Attention-driven dynamic graph convolutional network for multi-label image recognition,” in *Proc. European Conference on Computer Vision*, 2020, pp. 649–665.
- [28] Z. Chen, X.-S. Wei, P. Wang, and Y. Guo, “Learning graph convolutional networks for multi-label recognition and applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [29] J. Shi, W. Xiong, Z. Lin, and H. J. Jung, “Instantbooth: Personalized text-to-image generation without test-time finetuning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 8543–8552.
- [30] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, “Multi-concept customization of text-to-image diffusion,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1931–1941.
- [31] Y. Gu, X. Wang, J. Z. Wu, Y. Shi, Y. Chen, Z. Fan, W. Xiao, R. Zhao, S. Chang, W. Wu, et al., “Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models,” in *NeurIPS*, 2023.

- [32] Y. L. e. al., “Human attribute recognition by deep hierarchical contexts,” in *Proc. European Conf. Computer Vision*, 2016, p. 684–700.
- [33] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *Proc. International Conference on Learning Representations*, 2017.
- [34] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proc. IEEE International Conference on Machine Learning*, 2015, pp. 2048–2057.
- [35] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.
- [36] A. L. Maas, Awni, Hannun, and A. Y. Ng, “Rectifier nonlinearities improve neural network acoustic models,” in *Proc. International Conference on Machine Learning*, 2013.
- [37] T. L. e. al., “Microsoft coco: Common objects in context,” in *Proc. European Conf. Computer Vision*, 2014, pp. 740–755.
- [38] Q. Zheng, Z. Li, and A. Bargteil, “Learning to shadow hand-drawn sketches,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7436–7445.
- [39] K. Awane, K. Tsubota, H. Ikuta, Y. Matsui, K. Aizawa, and N. Yanase, “Improving the quality of illustrations: Transforming amateur illustrations to a professional standard,” in *Proc. IEEE International Conference on Image Processing*, 2021, pp. 584–588.
- [40] F. C. Silva, P. A. L. d. Castro, H. R. Júnior, and E. C. Marujo, “Mangan: Assisting colorization of manga characters concept art using conditional gan,” in *Proc. IEEE International Conference on Image Processing*, 2019, pp. 3257–3261.
- [41] T.-T. Fang, D. M. Vo, A. Sugimoto, and S.-H. Lai, “Stylized-colorization for line arts,” in *Proc. IEEE International Conference on Pattern Recognition*, 2021, pp. 2033–2040.
- [42] Y. Jin, J. Zhang, M. Li, Y. Tian, H. Zhu, and Z. Fang, “Towards the high-quality anime characters generation with generative adversarial networks,” in *Proc. NIPS Workshop on Machine Learning for Creativity and Design*, 2017.
- [43] B. Li, Y. Zhu, Y. Wang, C.-W. Lin, B. Ghanem, and L. Shen, “Anigan: Style-guided generative adversarial networks for unsupervised anime face generation,” *arXiv preprint arXiv:2102.12593*, 2021.

- [44] Z. Ge, D. Mahapatra, S. Sedai, R. Garnavi, and R. Chakravorty, “Chest X-rays classification: A multi-label and fine-grained problem,” *arXiv preprint arXiv:1807.07247*, 2018.
- [45] H. Jain, Y. Prabhu, and M. Varma, “Extreme multi-label loss functions for recommendation, tagging, ranking and other missing label applications,” in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 935–944.
- [46] P. Deng, J. Ren, S. Lv, J. Feng, and H. Kang, “Multi-label image recognition in anime illustration with graph convolutional networks,” in *Proc. AAAI Conference on Artificial Intelligence*, 2020.
- [47] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, “Panda: Pose aligned networks for deep attribute modeling,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1637–1644.
- [48] O. Russakovsky and F.-F. Li, “Attribute learning in large-scale datasets,” in *Proc. European Conference on Computer Vision*, 2010, pp. 1–14.
- [49] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang, “Learning hypergraph-regularized attribute predictors,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 409–417.
- [50] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.
- [51] X. Wang and J. Yu, “Learning to cartoonize using white-box cartoon representations,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8090–8099.
- [52] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, “Collecting image annotations using Amazon’s mechanical turk,” in *Proc. NAACL HLT Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, 2010, pp. 139–147.
- [53] G. Yang, N. Fei, M. Ding, G. Liu, Z. Lu, and T. Xiang, “L2M-GAN: Learning to manipulate latent space semantics for facial attribute editing,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2951–2960.
- [54] L. Zhang, C. Li, Y. Ji, C. Liu, and T. t. Wong, “Erasing appearance preservation in optimization-based smoothing,” in *Proc. European Conference on Computer Vision*, 2020.

- [55] J. Back, “Fine-tuning StyleGAN2 for cartoon face generation,” *arXiv preprint arXiv:2106.12445*, 2021.
- [56] J. Back, S. Kim, and N. Ahn, “Webtoonme: A data-centric approach for full-body portrait stylization,” *arXiv preprint arXiv:2210.10335*, 2022.
- [57] J. Lee, E. Kim, Y. Lee, D. Kim, J. Chang, and J. Choo, “Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2020.
- [58] L. Zhang, C. Li, E. Simo-Serra, Y. Ji, T.-T. Wong, and C. Liu, “User-guided line art flat filling with split filling mechanism,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2021.
- [59] S. Gopal and Y. Yang, “Recursive regularization for large-scale classification with hierarchical and graphical dependencies,” in *KDD*, 2013.
- [60] C. Fellbaum, “Wordnet: An electronic lexical database,” *Bradford Books*, 1998.
- [61] S. Gopal and Y. Yang, “Hierarchical bayesian inference and recursive regularization for large-scale classification,” *ACM Transactions on Knowledge Discovery from Data*, vol. 9, no. 3, 2015.
- [62] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *Proc. International Conference on Machine Learning*, 2015, pp. 2256–2265.
- [63] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv preprint arXiv:2204.06125*, 2022.
- [64] B. Kavar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, “Imagic: Text-based real image editing with diffusion models,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6007–6017.
- [65] L. Zhang and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” *arXiv preprint arXiv:2302.05543*, 2023.
- [66] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang, “Implicit diffusion models for continuous super-resolution,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10021–10030.

- [67] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, “Repaint: Inpainting using denoising diffusion probabilistic models,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11461–11471.
- [68] T. Brooks, A. Holynski, and A. A. Efros, “Instructpix2pix: Learning to follow image editing instructions,” *arXiv preprint arXiv:2211.09800*, 2023.
- [69] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao, “Nightshade: Prompt-specific poisoning attacks on text-to-image generative models,” in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024, pp. 807–825.
- [70] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *Proc. International Conference on Learning Representations*, 2018.
- [71] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.
- [72] A. Heng and H. Soh, “Selective amnesia: A continual learning approach to forgetting in deep generative models,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 17170–17194, 2023.
- [73] J. Vice, N. Akhtar, R. Hartley, and A. Mian, “BAGM: A backdoor attack for manipulating text-to-image generative models,” *IEEE Transactions on Information Forensics and Security*, 2024.
- [74] Y. Huang, Q. Guo, and F. Juefei-Xu, “Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models,” *arXiv preprint arXiv:2305.10701*, 2023.
- [75] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al., “Sam 2: Segment anything in images and videos,” *arXiv preprint arXiv:2408.00714*, 2024.
- [76] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.
- [77] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *Proc. IEEE Symposium on Security and Privacy*, 2017, pp. 39–57.
- [78] K. Gregor, F. Besse, D. Jimenez Rezende, I. Danihelka, and D. Wierstra, “Towards conceptual compression,” *Advances In Neural Information Processing Systems*, vol. 29, 2016.

- [79] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in neural information processing systems*, vol. 30, 2017.
- [80] M. Fréchet, “Sur la distance de deux lois de probabilité,” *Annales de l’ISUP*, vol. VI, no. 3, pp. 183–198, 1957.
- [81] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, “Assessing generative models via precision and recall,” *Advances in neural information processing systems*, vol. 31, 2018.
- [82] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [83] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
- [84] A. Horé and D. Ziou, “Image quality metrics: Psnr vs. ssim,” in *Proceedings of the International Conference on Pattern Recognition*, 2010, pp. 2366–2369.

Achievements of the Author

(A) Journal

- [A-1] Z. Lan, K. Maeda, T. Ogawa, and M. Haseyama, “Hierarchical multi-label attribute classification with graph convolutional networks on anime illustration,” *IEEE Access* (IF: 3.6), vol. 11, pp. 35447–35456, 2023.
- [A-2] Z. Lan, K. Maeda, T. Ogawa, and M. Haseyama, “Multi-label classification in anime illustrations based on hierarchical attribute relationships,” *Sensors* (IF: 3.5), vol. 23, no. 10, 4798, 2023.

(B) International Conference

- [B-1] Z. Lan, K. Maeda, T. Ogawa, and M. Haseyama, “Multi-label image recognition based on multi-modal graph convolutional networks using captioning features,” in *Proceedings of the IEEE Global Conference on Consumer Electronics (GCCE)*, pp. 273–274, 2021.
- [B-2] Z. Lan, K. Maeda, T. Ogawa, and M. Haseyama, “GCN-based multi-modal multi-label attribute classification in anime illustration using domain-specific semantic features,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 2021–2025, 2022.
- [B-3] Z. Lan, K. Maeda, T. Ogawa, and M. Haseyama, “Adversarial perturbation shield: preventing concept bleed-through in continual learning of personalized generative models,” in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026. (Accept)

(C) Domestic Conference

- [C-1] 蘭子文, 前田 圭介, 小川 貴弘, 長谷山 美紀, “グラフ畳み込みネットワークに基づくアニメイラストのマルチラベル画像認識に関する検討,” 令和3年度電気・情報関係学会 北海道支部連合大会, pp. 125–126, 2021.
- [C-2] Z. Lan, K. Maeda, T. Ogawa, and M. Haseyama, “A note on multi-label image recognition in anime illustration based on graph convolutional networks using captioning features,” ITE Technical Report, vol. 46, no. 6, pp. 161–165, 2022.

[C-3] Z. Lan, K. Maeda, T. Ogawa, and M. Haseyama, “A note on multi-label image classification in animation illustration considering hierarchical relationships of attributes,” ITE Technical Report, vol. 47, no. 6, pp. 212–216, 2023.

[C-4] Z. Lan, K. Maeda, T. Ogawa, and M. Haseyama, “Adversarial attack focused on manipulating semantic information for personalized text-to-image diffusion models,” Meeting on Image Recognition and Understanding (MIRU), pp. 1–5, 2024.

[C-5] Z. Lan, K. Maeda, T. Ogawa, and M. Haseyama, “Adversarial perturbation to prevent concept bleed-through in continual learning of personalized generative models,” Meeting on Image Recognition and Understanding (MIRU), pp. 1–5, 2025.

(D) Award

[D-1] IEEE GCCE 2021 Outstanding Paper Award, 2021.

[D-2] The 2024 IEEE Sapporo Section Encouragement Award, 2025.

(E) Scholarship

[E-1] Hokkaido University Ambitious Doctoral Fellowship, 2023.4-2024.3.

[E-2] Hokkaido University EXEX Doctoral Fellowship, 2024.4-2024.10.

[E-3] Hokkaido University Next Generation AI Doctoral Fellowship, 2024.10-2026.3.