Wrocław University of Science and Technology
Faculty of Information and Communication Technology

Field of study: **Artificial Intelligence**
Speciality: **---**

MASTER THESIS

A Method for Image Generation Using Conditional Multi-Views

Eryk Wójcik

Supervisor
dr hab. inż. Maciej Zięba

Machine Learning, Generative Models, Diffusion

WROCŁAW 2025

Streszczenie

Dodaj streszczenie pracy w języku polskim. Staraj się uwzględnić wymienione na stronie tytułowej słowa kluczowe. Uwaga przedstawiony rekomendowany szablon dotyczy pracy dyplomowej pisanej w języku angielskim. W przeciwnym wypadku, student powinien samodzielnie zmienić nazwy „Chapter" na „Rodział" itp stosując odpowiednie pakiety systemu LaTeXoraz ustawienia w pliku *latex-settings.tex*.

Abstract

Streszczenie w języku angielskim.

Spis treści

1 Introduction TODO **1**
1.1 Problem Statement 1
1.2 Thesis Objectives 1
1.3 Thesis Outline 1

2 Related Work **3**
2.1 SLAM-based Solutions for Novel View Synthesis 3
2.2 Image-Prompt Based Generative Models 3
2.2.1 Text-to-Image and Image-to-Image Models 4
2.2.2 Multi-view Diffusion Models 4
2.3 Adapter-based Methods for Multi-view Generation 4
2.3.1 Adapter Mechanisms in Diffusion Models 4
2.3.2 Multi-view Adapters 5
2.4 Limitations and Research Gaps 5

3 Conclusions and Future Work TODO **7**

1. Introduction TODO

W pracy formułuje się cele o charakterze badawczym wymagające doboru i zastosowania metod badawczych, wykorzystując wiedzę teoretyczną oraz naukową. Wskazane jest przedstawienie, co nowego jest zaproponowane w pracy oraz podanie ograniczeń i słabych/mocnych stron opracowanego rozwiązania (jeżeli dotyczy). Rozdział wprowadzający powinien służyć czytelnikowi do zrozumienia celu pracy.

1.1. Problem Statement

W tej sekcji student powinien przedstawić bliżej problem, którym chce się zmierzyć. Jasno zdefiniuj problem badawczy. Podaj swoje cele, zadania i pytania badawcze. Wyjaśnij znaczenie badania. Określ ograniczenia badań.

1.2. Thesis Objectives

W tej sekcji powinny zostać przedstawione konkretne działania, które określają pracę studenta w celu rozwiązania problemu.

1.3. Thesis Outline

Zarysuj strukturę swojej pracy dyplomowej. Ogólnie przedstawienie pracy. Przykładowo: „Praca dzieli się na 7 rozdziałów (…)". Rozdział **??** dotyczy (…). Temat został rozwinięty w **??**.

2. Related Work

In this chapter, I review existing approaches to multi-view image generation and novel view synthesis. Starting with traditional SLAM-based methods for novel view synthesis (Section 2.1), followed by prompt-image based generative models for view synthesis (Section 2.2), with particular emphasis on adapter-based methods to fine-tune diffusion models to generate views conditioned on a single image (Section 2.3). Finally, I analyze the limitations of current methods and identify research gaps that my work aims to address (Section 2.4).

2.1. SLAM-based Solutions for Novel View Synthesis

Traditional approaches to novel view synthesis have relied heavily on Structure from Motion (SfM) and Simultaneous Localization and Mapping (SLAM) techniques. These methods focus on reconstructing 3D geometry from multiple images and then rendering novel views based on this geometry.

Early works in this domain, such as those described by Hartley and Zisserman [3], established the mathematical foundations for structure from motion algorithms. These approaches typically involve feature extraction (using methods like SIFT [5], ORB [10], etc.), matching across images, camera pose estimation, and 3D reconstruction through triangulation. While these methods provide geometrically accurate reconstructions, they often struggle with texturing and rendering photorealistic novel views, especially in regions with limited observations or complex materials. These methods are not able to handle sparse inputs (e.g., a single image or a few images) and require dense multi-view captures (often hundreds of images) to create high-quality 3D representations. Another limitation is the time it takes to run these algorithms, especially if we match features across many high resolution images.

More recent advancements in Neural Radiance Fields (NeRF) [7] have significantly improved the quality of novel view synthesis by representing scenes as continuous volumetric functions that map 3D coordinates and viewing directions to color and density values. However, NeRF and its variants typically require dense multi-view captures (often hundreds of images) to create high-quality 3D representations, limiting their practical applicability in scenarios where only a few images are available.

2.2. Image-Prompt Based Generative Models

The emergence of powerful diffusion models has revolutionized the field of image generation, including novel view synthesis. These models have demonstrated remarkable capabilities in generating high-quality images conditioned on various inputs, such as text prompts, reference images, or camera poses.

2.2.1. Text-to-Image and Image-to-Image Models

Text-to-image diffusion models like Stable Diffusion [9] have shown impressive capabilities in generating diverse and high-quality images from textual descriptions. Building upon these foundations, several works have extended these models to handle image-to-image translation tasks, where a reference image serves as an additional conditioning signal.

Zero-1-to-3 [6] pioneered the approach of conditioning diffusion models on both a reference image and camera pose information to generate novel views. This method demonstrated the potential of leveraging pre-trained text-to-image models for novel view synthesis without requiring explicit 3D reconstruction. However, it often struggles with maintaining geometric consistency across generated views.

2.2.2. Multi-view Diffusion Models

To address the limitations of single-view approaches, several works have focused on developing multi-view diffusion models that can generate multiple consistent views simultaneously.

MVDream [11] extends the self-attention mechanism in diffusion models to operate across multiple views, enabling the generation of 3D-consistent images. By jointly modeling multiple views, this approach significantly improves geometric consistency compared to methods that generate each view independently.

Similarly, ViewCrafter [12] combines video latent diffusion models [1] with 3D point cloud priors to generate high-fidelity and consistent novel views. By leveraging the explicit 3D information provided by point clouds and the generative capabilities of video diffusion models, ViewCrafter achieves precise control of camera poses and generates high-quality novel views.

CAT3D [2] takes a different approach by simulating a real-world capture process with a multi-view diffusion model. Given one or three input images and a set of target novel viewpoints, this model generates highly consistent novel views that can be used as input to robust 3D reconstruction techniques.

While these multi-view diffusion models have shown impressive results, they typically require full fine-tuning of pre-trained text-to-image models, which is computationally expensive and may lead to degradation in image quality due to the scarcity of high-quality 3D data.

2.3. Adapter-based Methods for Multi-view Generation

To address the limitations of full fine-tuning approaches, recent works have explored adapter-based methods that allow for more efficient adaptation of pre-trained models to specific tasks while preserving their original capabilities.

2.3.1. Adapter Mechanisms in Diffusion Models

Adapters are lightweight modules that can be inserted into pre-trained models to adapt them to new tasks without modifying the original network parameters. This approach has gained popularity in natural language processing and has also been applied to diffusion models for various image generation tasks.

ControlNet [13] introduced a method to add spatial conditioning to text-to-image diffusion models by training additional control modules that are connected to the original UNet backbone. This approach allows for precise control over the generated images while preserving the original model's capabilities.

Similarly, T2I-Adapter [8] proposed a more modular approach where adapters are trained separately and can be combined to provide multiple forms of control simultaneously. These methods have demonstrated the effectiveness of adapter-based approaches for controlled image generation.

2.3.2. Multi-view Adapters

Building upon the success of adapter mechanisms, MV-Adapter [4] introduced the first adapter-based solution for multi-view image generation. Unlike previous approaches that make invasive modifications to pre-trained text-to-image models and require full fine-tuning, MV-Adapter enhances these models with a plug-and-play adapter that preserves the original network structure and feature space.

MV-Adapter employs a decoupled attention mechanism, where the original spatial self-attention layers are retained, and new multi-view attention layers are created by duplicating the structure and weights of the original layers. These layers are organized in a parallel architecture, allowing the adapter to inherit the powerful priors of the pre-trained self-attention layers while efficiently learning geometric knowledge.

Additionally, MV-Adapter introduces a unified condition encoder that seamlessly integrates camera parameters and geometric information, facilitating applications such as text and image-based 3D generation and texturing. By updating fewer parameters, MV-Adapter enables efficient training and preserves the prior knowledge embedded in pre-trained models, mitigating overfitting risks.

2.4. Limitations and Research Gaps

Despite the significant progress in multi-view image generation and novel view synthesis, several limitations and research gaps remain:

1. **Computational Efficiency**: Full fine-tuning of diffusion models for multi-view generation is computationally expensive, especially when working with large base models and high-resolution images. While first adapter-based method MV-Adapter has improved efficiency of training, there is still room for improvement.

2. **Geometric Consistency**: Maintaining geometric consistency across generated views remains a challenge, particularly when generating views from significantly different perspectives. Current methods often struggle with complex occlusions, reflective surfaces and fine geometric details.

3. **Sparse Input Handling**: Most existing methods require either dense multi-view captures or make strong assumptions about the scene structure. There is a need for methods that can effectively handle sparse inputs (e.g., a single image or a few images) while generating high-quality novel views.

4. **Integration of Geometric Priors**: While some methods incorporate geometric information through camera poses or point clouds, the effective integration of these priors with generative models remains an open research question.

My work aims to address these limitations by developing a method that combines the efficiency of adapter-based approaches with the geometric consistency provided by point cloud priors. Specifically, I propose to extend the MV-Adapter framework by incorporating point cloud information as an additional conditioning signal, similar to the approach used in ControlNet. This will allow for more precise control over the generated views while maintaining the computational efficiency of adapter-based methods.

3. Conclusions and Future Work TODO

Zakończenie, podsumowuje najważniejsze wnioski, podaje możliwości dalszego rozwinięcia wykonanych prac i wskazuje obszar potencjalnego zastosowania pracy. Rezultaty pracy mają charakter poznawczy, mogą mieć charakter użytkowy. Należy dokonać analizy uzyskanych wyników. Rezultaty powinny charakteryzować się oryginalnością, a nawet w pewnym stopniu nowatorstwem. Praca zawiera (…). Zostało pokazane (…). Eksperymenty wykazały (…). Tu piszemy wnioski i obserwacje.

Widzimy, że (…). Z tego powodu przyszła praca powinna obejmować (…).

Bibliografia

[1] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis. Align your latents: High-resolution video synthesis with latent diffusion models, 2023.

[2] R. Gao, A. Holynski, P. Henzler, A. Brussee, R. Martin-Brualla, P. Srinivasan, J. T. Barron, and B. Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024.

[3] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003.

[4] Z. Huang, Y. Guo, H. Wang, R. Yi, L. Ma, Y.-P. Cao, and L. Sheng. Mv-adapter: Multi-view consistent image generation made easy. *arXiv preprint arXiv:2412.03632*, 2024.

[5] T. Lindeberg. Scale invariant feature transform. *Scholarpedia*, 7, 05 2012.

[6] R. Liu, R. Wu, B. V. Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023.

[7] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020.

[8] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[10] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: an efficient alternative to sift or surf. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2564–2571, 11 2011.

[11] Y. Shi, P. Wang, J. Ye, L. Mai, K. Li, and X. Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023.

[12] W. Yu, J. Xing, L. Yuan, W. Hu, X. Li, Z. Huang, X. Gao, T.-T. Wong, Y. Shan, and Y. Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis, 2024.

[13] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

Spis rysunków

Spis tabel