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

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.

2. Related Work

In this chapter, I provide a comprehensive review of existing approaches relevant to multi-view image generation.

The chapter begins by introducing the fundamental task of Novel View Synthesis (NVS) and its significance. We then delve into traditional 3D reconstruction techniques (Section 2.1) and their inherent limitations, particularly for sparse-input scenarios, which motivates the exploration of generative methods. Subsequently, the discussion shifts to the foundational principles of Deep Generative Models for Image Synthesis, with a focus on diffusion models (Section 2.2).

Building on this, we explore various methods for Conditioning Diffusion Models for Enhanced Control and New Tasks (Section 2.3), including the crucial role of camera parameter encoding and the concept of lightweight adaptation through adapters.

The core of the chapter then examines state-of-the-art Diffusion-based Multi-View Image Generation techniques (Section 2.4), covering both single reference image novel view synthesis and architectures for coherent multi-view generation, including specialized multi-view adapters.

2.1. Traditional 3D Reconstruction Approaches

Simultaneous Localization and Mapping (SLAM) is a fundamental concept in the field of computer vision and robotics. It refers to the process of simultaneously estimating the camera's position and orientation in a 3D environment while mapping the environment itself. Originally, SLAM was used to track the position of a robot in a 3D space, but it has since been applied to a wide range of problems, including augmented reality, medical applications and novel view synthesis.

SLAM works by processing sensor data in real-time to create a map of the unknown environment while simultaneously tracking the position of the sensor within that map. This process typically involves several key steps:

1. **Feature Detection**: Identifying distinctive points or features in the environment from sensor data
2. **Data Association**: Matching observed features with previously mapped features
3. **State Estimation**: Updating the estimated pose of the camera and the map of the environment
4. **Loop Closure**: Recognizing when the sensor has returned to a previously visited location and adjusting the map accordingly to reduce accumulated errors

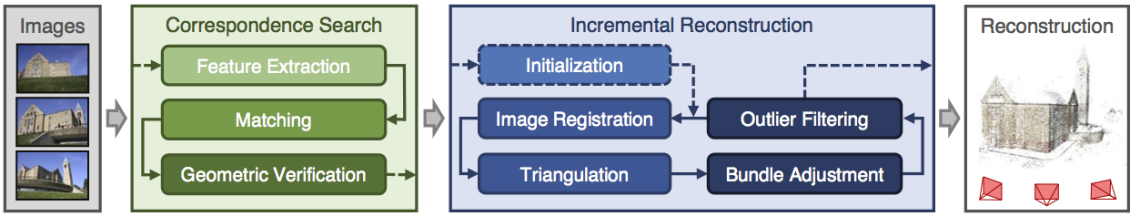
To address these requirements, researchers have concentrated on creating techniques for machines to independently build ever more precise scene representations. This integration of

robotics, computer vision, sensor technology, and recent advancements in artificial intelligence has shaped this field.

Typically, SLAM techniques use a combination of data sources, such as images, laser range scans, sonars and GPS to effectively map an environment. In this work I will focus on the use of images for 3D reconstruction.

2.1.1. Structure from Motion Pipeline

One of the most popular methods for 3D reconstruction from images is COLMAP [14]. COLMAP is a general-purpose structure-from-motion (SfM) [5] system that can automatically reconstruct 3D scenes from a collection of images. It is a popular choice for 3D reconstruction due to its accuracy, speed, and ease of use.



Rysunek 2.1: COLMAP pipeline

Structure from Motion works in two main steps [Figure 2.1]:

1. **Correspondence Search**: Identify the unique landmarks (features) in all of the images and match the same landmarks across images.

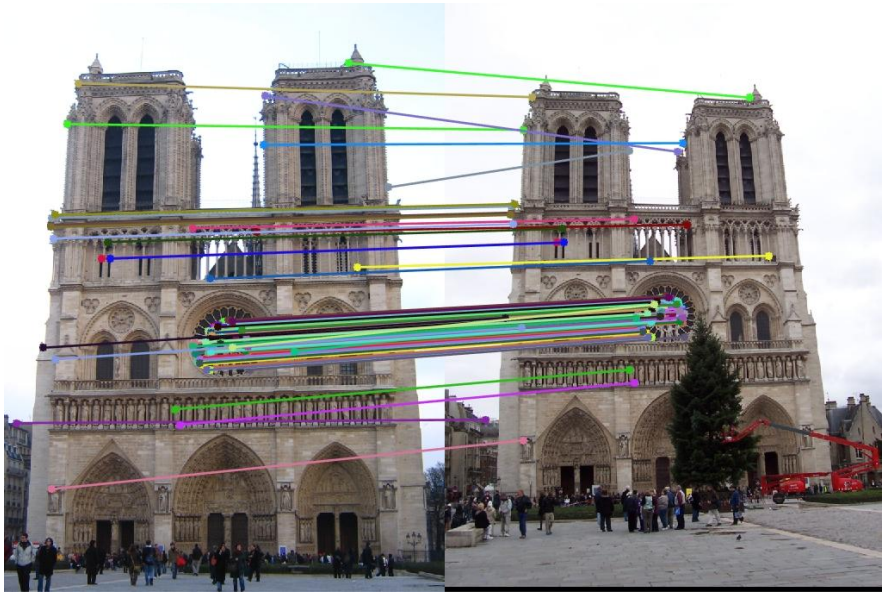
2. **Incremental Reconstruction**: Estimates the camera poses and triangulates 3D points through an iterative process.

First step of correspondence search is feature extraction. COLMAP uses SIFT [7] and ORB [13] methods to extract features from images. Scale-invariant feature transform (SIFT) is a method for detecting keypoints that contrast with their surroundings and describing the local image content around them. It is invariant to rotation and scale, making it robust for matching features across images taken from different viewpoints and distances. Oriented FAST and Rotated BRIEF (ORB) is a more computationally efficient alternative to SIFT, combining a high-speed FAST detector with optimized descriptors for rotation invariance, offering comparable matching performance with significantly reduced computational requirements.

Second step is matching features across images. Feature matching is a process of finding the best matches of features across the images. COLMAP uses a variant of the FLANN [11] library to find the best matches for each feature. FLANN is a library for performing fast approximate nearest neighbor searches in high dimensional spaces. Feature matching is visualized in Figure 2.2.

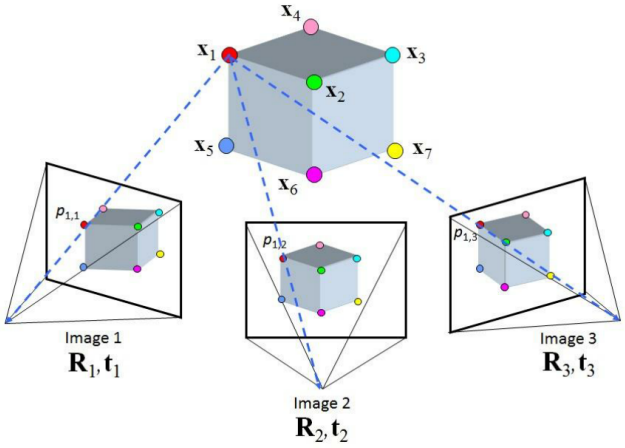
Then the geometric verification step uses the prior knowledge about the camera model and motion to remove outliers, preparing the verified feature matches for the subsequent Incremental Reconstruction phase.

When it comes to Incremental Reconstruction, the process is as follows:



Rysunek 2.2: Feature matching result

1. **Camera Pose Estimation**: Estimate the camera location and direction in 3D space for each image.
2. **Triangulation**: Triangulate 3D points of the observed objects from the camera poses and matched features.
3. **Bundle Adjustment**: Refine the camera poses and 3D points to minimize the reprojection error.



Rysunek 2.3: Camera pose estimation

The camera pose estimation (visualized in Figure 2.3) starts with an initialization step, where the initial pair of images and matched features between them are used to estimate the camera pose. Then, the algorithm proceeds to the loop (Figure 2.1), where the next image

is registered to the reconstruction and the camera pose is estimated again (triangulation step). After that the bundle adjustment step is performed to refine the camera poses and 3D points to be consistent with the entire dataset with images of a given scene. This optimization process typically uses the Levenberg-Marquardt algorithm to minimize reprojection error. This process is repeated for each subsequent batch of images, updating the camera poses and 3D points.

2.1.2. Methods for 3D Reconstruction

The field of 3D reconstruction encompasses various approaches beyond the basic Structure from Motion pipeline. These methods can be broadly categorized into two groups: sparse reconstruction (like SfM) and dense reconstruction methods.

Dense Reconstruction Methods

While SfM provides camera poses and a sparse point cloud, dense reconstruction methods aim to create a complete 3D model with detailed surface information. Notable methods include:

1. **Multi-View Stereo (MVS)**: After obtaining camera poses through SfM, MVS algorithms like PMVS [2] generate dense point clouds by matching pixels across multiple images.

2. **Depth Map Fusion**: Methods such as COLMAP's MVS pipeline estimate per-image depth maps and then fuse them into a consistent 3D model.

3. **Neural Radiance Fields (NeRF)** [9]: A more recent approach that represents scenes as continuous 5D functions (spatial location and viewing direction) encoded in neural networks. NeRF takes camera poses from SfM as input and uses ray tracing to synthesize novel views with remarkable detail and view consistency.

Learning-based Reconstruction

Recent approaches leverage deep learning for 3D reconstruction, often using SfM-derived data for supervision or initialization:

1. **Learned MVS**: Methods like MVSNet [17] use convolutional neural networks to learn the depth estimation process directly from images and camera parameters.

2. **Single-view Reconstruction**: Networks like Mesh R-CNN [4] can estimate 3D structure from a single image by leveraging prior knowledge learned from large datasets.

2.1.3. Limitations of 3D Reconstruction

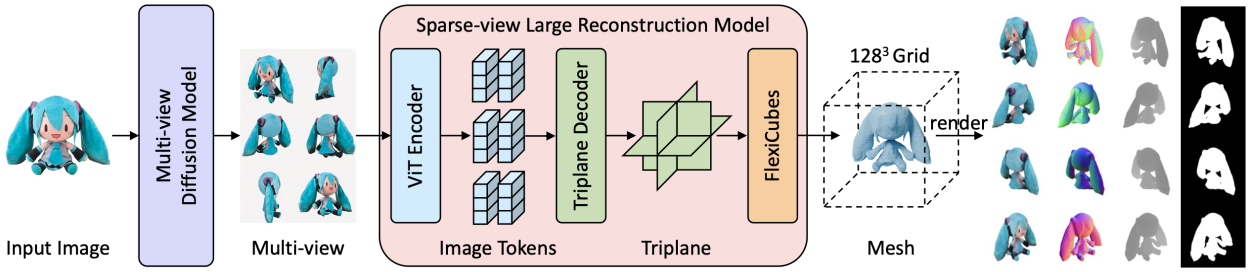
Structure from Motion is a powerful tool for 3D reconstruction, demonstrating high effectiveness across a variety of scenarios. However, it encounters significant challenges. These include dealing with textureless surfaces, reflective materials, and the computational

complexity of processing high-resolution images. Most importantly in the context of this thesis, it struggles with sparse input scenarios, such as those involving a single image or only a few images.

Traditional 3D reconstruction methods like SfM and MVS typically require a dense collection of images with sufficient overlap to establish accurate feature correspondences and camera pose estimations. When faced with limited input views—particularly in the extreme case of a single image—these methods often fail to generate complete and accurate 3D representations. The quality of reconstruction degrades significantly due to:

1. **Geometric ambiguity**: A single image or sparse set of images provides incomplete information about occluded regions and depth, leading to ambiguous geometry.
2. **Feature matching limitations**: Fewer images means fewer opportunities to establish reliable feature correspondences across different viewpoints.
3. **Inability to triangulate**: Robust triangulation requires features to be visible from multiple viewpoints, which is not possible with very limited inputs.
4. **View-dependent effects**: Materials with specular reflections or varying appearance based on viewpoint cannot be accurately modeled without multiple observations.

These limitations have motivated the development of generative approaches to novel view synthesis, particularly using diffusion models trained on large datasets of rendered images of 3D models. Instead of explicitly reconstructing geometry, these methods leverage the power of deep learning to hallucinate plausible views from unseen perspectives.



Rysunek 2.4: InstantMesh

Modern approaches of 3D object generation, such as InstantMesh [16], leverage these diffusion models to generate multiple consistent views of an object from minimal input (a single image or even a text prompt). This novel view synthesis is the first step in a pipeline 2.4 that can be used to create a 3D model of the object, and it is this particular step that forms the central focus of this thesis.

2.2. Deep Generative Models for Image Synthesis

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. Diffusion Models

UNET

2.2.2. Text-to-Image Diffusion Models

Text-to-image diffusion models like Stable Diffusion [12] 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.

2.3. Conditioning Diffusion Models for Enhanced Control and New Tasks

2.3.1. Diverse Conditioning Signals

2.3.2. Camera Parameter Encoding for 3D Awareness

2.3.3. Lightweight Adaptation

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.

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 [19] 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 [10] 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.

2.4. Diffusion-based Multi-View Image Generation

2.4.1. Single Reference Image Novel View Synthesis

Zero-1-to-3 [8] 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.

2.4.2. Coherent Multi-View Generation Architectures

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 [15] 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 [18] 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 [3] 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.

2.4.3. Specialized Multi-View Adapters

Building upon the success of adapter mechanisms, MV-Adapter [6] 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.

problems of the current MVD: - inconsistant lighting - poor geometric consistency - big model training requirements

3. Proposed Method TODO

In this chapter, I describe the proposed method.

3.1. Overview

3.2. Limitations of Previous Methods

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.

3.3. Multi-View Image Generation

3.4. Conditional Multi-View Image Generation

4. Data Preparation TODO

4.1. Datasets

4.1.1. Synthetic Datasets

4.1.2. Real Datasets

4.2. Data Augmentation

5. Experiments TODO

In this chapter, I describe the experiments I conducted to evaluate the proposed method.

6. Results TODO

In this chapter, I present the results of the experiments.

7. 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] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010.

[3] 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.

[4] G. Gkioxari, J. Malik, and J. Johnson. Mesh r-cnn, 2020.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] M. Muja and D. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *FAST APPROXIMATE NEAREST NEIGHBORS WITH AUTOMATIC ALGORITHM CONFIGURATIONs*, volume 1, pages 331–340, 01 2009.

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

[13] 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.

[14] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

[16] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

[17] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[18] 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.

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

Spis rysunków

Spis tabel