Department of Computer Science

Learning Latent Image Representations with Prior Knowledge

Yuxin Hou

Aalto University

DOCTORAL THESES

Aalto University publication series
DOCTORAL THESES 194/2022

Learning Latent Image Representations with Prior Knowledge

Yuxin Hou

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall T2 of the school on 16 December 2022 at 13:15.

Aalto University
School of Science
Department of Computer Science

Supervising professors
Prof. Juho Kannala, Aalto University, Finland,
Prof. Arno Solin, Aalto University, Finland

Preliminary examiners
Prof. Oisin Mac Aodha, University of Edinburgh, UK
Prof. Martin R. Oswald, University of Amsterdam, Netherlands

Opponent
Prof. Oisin Mac Aodha, University of Edinburgh, UK

Aalto University publication series
DOCTORAL THESES 194/2022

ISBN 978-952-64-1072-2 (printed)
ISBN 978-952-64-1073-9 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-1073-9

Unigrafia Oy
Helsinki 2022

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Yuxin Hou

Name of the doctoral thesis
Learning Latent Image Representations with Prior Knowledge

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL THESES 194/2022

Field of research Computer Science

Manuscript submitted 7 June 2022 **Date of the defence** 16 December 2022

Permission for public defence granted (date) 24 October 2022 **Language** English

☐ **Monograph** ☒ **Article thesis** ☐ **Essay thesis**

Abstract

Deep learning has become a dominant tool in many computer vision applications due to the superior performance of extracting low-dimensional latent representations from images. However, though there is prior knowledge for many applications already, most existing methods learn image representations from large-scale training data in a black-box way, which is not good for interpretability and controllability.

This thesis explores approaches that integrate different types of prior knowledge into deep neural networks. Instead of learning image representations from scratch, leveraging the prior knowledge in latent space can softly regularize the training and obtain more controllable representations.The models presented in the thesis mainly address three different problems:
(i) How to encode epipolar geometry in deep learning architectures for multi-view stereo. The key of multi-view stereo is to find the matched correspondence across images. In this thesis, a learning-based method inspired by the classical plane sweep algorithm is studied. The method aims to improve the correspondence matching in two parts: obtaining better potential correspondence candidates with a novel plane sampling strategy and learning the multiplane representations instead of using hand-crafted cost metrics.
(ii) How to capture the correlations of input data in the latent space. Multiple methods that introduce Gaussian process in the latent space to encode view priors are explored in the thesis. According to the availability of relative motion of frames, there is a hierarchy of three covariance functions which are presented as Gaussian process priors, and the correlated latent representations can be obtained via latent nonparametric fusion. Experimental results show that the correlated representations lead to more temporally consistent predictions for depth estimation, and they can also be applied to generative models to synthesize images in new views.
(iii) How to use the known factors of variation to learn disentangled representations. Both equivariant representations and factorized representations are studied for novel view synthesis and interactive fashion retrieval respectively.

In summary, this thesis presents three different types of solutions that use prior domain knowledge to learn more powerful image representations. For depth estimation, the presented methods integrate the multi-view geometry into the deep neural network. For image sequences, the correlated representations obtained from inter-frame reasoning make more consistent and stable predictions. The disentangled representations provide explicit flexible control over specific known factors of variation.

Keywords Deep Learning, Machine Learning, Computer Vision, Multi View Stereo, Novel View Synthesis, Gaussian Processes

ISBN (printed) 978-952-64-1072-2 **ISBN (pdf)** 978-952-64-1073-9

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2022

Pages 135 **urn** http://urn.fi/URN:ISBN:978-952-64-1073-9

Preface

The work presented in the thesis has been carried out at the Department of Computer Science at Aalto University during years 2018-2022, and the research internship at Amazon Berlin from September 2020 to February 2021.

Firstly, I would like to express my greatest gratitude to my supervisors Prof. Juho Kannala and Prof. Arno Solin who provided me with resources to pursue my research interests and gave me advises throughout all these years. Thank you for your support and guidance. I am also grateful to the reviewers of the thesis, Dr. Oisin Mac Aodha from the University of Edinburgh and Dr. Martin R. Oswald from the University of Amsterdam for pre-examining the thesis and providing valuable comments.

During my years at Aalto University, I have had opportunities to work with many talented individuals at the Department of Computer Science. We closely share research ideas and learn from each other. I would like to thank a number of my lab mates in alphabetical order: Rinu Boney, Santiago Cortes, Ari Heljakka, Xiaotian Li, Iaroslav Melekhov, Shuzhe Wang and Rongtian Ye.

I have also had two great research internship opportunities during my doctoral studies. I would like to thank Dr. Loris Bazzani and Dr. Eleonora Vig for being my mentors during my stay at Amazon Berlin. I am also grateful to Dr. Tianwei Shen, Dr. Tsun-Yi Yang, and other team members from Facebook Reality Labs for all the inspiring discussions and collaborations.

Last but not least, I would like to thank my parents Xiaoming Hou and Shihui Pan for their unconditional love and support. And I want to thank my boyfriend, Yi Zhao, for his company during the past years.

Helsinki, November 25, 2022,

Yuxin Hou

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Yuxin Hou, Arno Solin and Juho Kannala. Unstructured Multi-view Depth Estimation Using Mask-Based Multiplane Representation. In *Scandinavian Conference on Image Analysis (SCIA)*, Norrköping, Sweden, pp. 54-66, June 2019.

II Yuxin Hou, Juho Kannala and Arno Solin. Multi-View Stereo by Temporal Nonparametric Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, pp. 2651-2660, October 2019.

III Yuxin Hou, Muhammad Kamran Janjua, Juho Kannala and Arno Solin. Movement-induced Priors for Deep Stereo. In *International Conference on Pattern Recognition (ICPR)*, Virtual, pp. 3628-3635, January 2021.

IV Yuxin Hou, Ari Heljakka and Arno Solin. Gaussian Process Priors for View-Aware Inference. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Virtual, pp. 7762-7770, May 2021.

V Yuxin Hou, Arno Solin and Juho Kannala. Novel View Synthesis via Depth-guided Skip Connections. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Virtual, pp. 3119-3128, January 2021.

VI Yuxin Hou, Eleonora Vig, Michael Donoser and Loris Bazzani. Learning Attribute-driven Disentangled Representations for Interactive Fashion Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Virtual, pp. 12147-12157, October 2021.

Author's Contribution

Publication I: "Unstructured Multi-view Depth Estimation Using Mask-Based Multiplane Representation"

Hou proposed the original idea and had the main responsibility in running all experiments and writing the article. Kannala and Solin gave valuable suggestions to improve the methods and contributed to writing the article.

Publication II: "Multi-View Stereo by Temporal Nonparametric Fusion"

Hou had the main responsibility in implementing the methods, running all experiments and writing the article. Solin proposed the original idea and provide reference code in Matlab for implementing Sec. 3.4. Kannala and Solin contributed to writing the article.

Publication III: "Movement-induced Priors for Deep Stereo"

Hou had the main responsibility in implementing the methods and running experiments. Janjua helped in running the experiments with the ZED dataset. Solin had the original idea. Solin and Kannala reviewed and proposed suggestions to the manuscript.

Publication IV: "Gaussian Process Priors for View-Aware Inference"

Hou implemented the methods used in Sec. 4.1, ran the related experiments and contributed to writing the article. Heljakka implemented the methods used in Sec. 4.2 and wrote related sections. Solin had the original idea and wrote the most of Sec. 1–3 and 5.

Publication V: "Novel View Synthesis via Depth-guided Skip Connections"

Hou proposed the original idea and implemented all the models and methods used in the article as well as evaluated the experiments. Solin and Kannala gave valuable comments during the discussions about the methods, results and writing.

Publication VI: "Learning Attribute-driven Disentangled Representations for Interactive Fashion Retrieval"

Hou proposed the original idea and had the main responsibility in designing the methods and conducting all experiments. Vig and Bazzani gave many valuable comments to improve the methods and experiments. All co-authors contributed to writing the article.

Abbreviations

AI Artificial Intelligence

CNN Convolutional Neural Network

GAN Generative Adversarial Network

GP Gaussian Process

IBR Image-based Rendering

MVS Multi View Stereo

NVS Novel View Synthesis

VAE Variational Autoencoder

1. Introduction

Computer vision is essential for developing artificial intelligence systems as it enables machines to understand the observed environment. It aims to extract information to describe the world from image data. Today, computer vision has been used in a wide variety of real-world applications like autonomous driving, robot navigation, image retrieval, *etc.*

In recent years, deep learning has become the dominant tool for computer vision because of the ability to learn complex low-dimensional latent representations of high-dimensional data. Instead of using hand-crafted features, deep learning methods learn latent representations from input images via a number of non-linear layers, and the learned representations have achieved better performance in different tasks like image recognition, semantic segmentation, *etc.* Moreover, learned representations enable AI systems to adapt to new tasks quickly compared to hand-crafted features. Though deep learning shows the promising potential of learning latent representations from visual data, it is not good for interpretability. On the other hand, since we have prior knowledge of the world already, we can utilize it to learn better latent representations. This thesis explores some ideas to encode different types of prior knowledge in the latent space for deep learning methods.

Basic mathematics research gives a comprehensive understanding of geometric relationships between multiple views [26], In that case, some complex geometric relationships do not need to be learned from scratch by deep neural networks. For example, given a pair of images from rectified stereo cameras, the depth can be observed from stereo disparity. Many traditional 3D vision methods are based on multi-view geometry. Though these classical methods have shown good results under most ideal Lambertian scenes, they are not robust enough. For example, challenging regions like texture-less regions or reflective surfaces are often difficult to be reconstructed. Since deep learning methods can introduce global semantic information, the integration of multi-view geometry and deep neural networks shows great potential in improving the prediction quality. Chapter 2 will review the learning-based methods for multi-view depth

estimation.

While frame-independent predictions with deep neural networks have become the dominant solution to many computer vision tasks, in some applications, the images are strongly correlated (such as the temporally consecutive camera frames), and the potential benefits of inter-frame reasoning have received less attention. On the other hand, probabilistic machine learning provides the ability to encode correlation as prior knowledge for inference. Chapter 3 will present methods that combine Gaussian processes and deep neural networks for different computer vision tasks like multi-view depth estimation, stereo matching and novel view synthesis.

Since computer vision aims to extract useful information to understand the world from images, it is beneficial to learn latent representations that can identify and disentangle the underlying explanatory factors. For some applications, the underlying factors are well-studied and we can leverage the specific domain knowledge of data. For example, for images captured by a camera, we know they depend on both the observed scene, camera model and camera pose; for images of clothing items, we know the semantic attributes like color, shape, and length influence the appearances. This domain knowledge can be used to help design algorithms for learning representations that can disentangle the factor of variation. Chapter 4 will present methods that utilize the prior knowledge of data domain to learn disentangled representations for computer vision tasks like novel view synthesis and image retrieval.

1.1 Contributions

All publications and codes are available as open access. The main contributions of the thesis are listed below:

- In Publication I, a method for multi-view depth estimation is introduced. Inspired by the traditional plane sweep algorithm, the proposed framework is the integration of multi-view geometry and deep neural networks. A novel plane sampling strategy is proposed to provide better correspondence candidates. The mask-based multiplane representations are learned to aggregate geometry information without using hand-crafted cost metrics after building the volume via differentiable homography warping.

- Publication II, Publication III and Publication IV explore the combination of probabilistic methods and deep neural networks to learn correlated representations for different 3D vision tasks. For images with given camera poses, Publication II proposes a soft constraint in latent space via a tailored Gaussian process prior to alleviate the temporal

inconsistency in multi-view depth estimation. Publication III extends Publication II to the stereo matching problem and proposes a new prior kernel that is based on gyroscope data. Publication IV proposes a novel kernel to encode 3D camera orientation. The proposed kernel can be used to manipulate the latent representations in generative models to synthesize novel view images.

- Publication V presents a method that learns equivariant representations for novel view synthesis, which takes advantages of both image-based rendering methods and pixel generation methods. By leveraging the transforming auto-encoders [31], the learned equivariant representations can predict depth maps for novel target views, and the predicted depths guide the alignment of the feature maps for skip connections.

- Publication VI presents a method that uses known semantic visual attributes to learn attribute-specific disentangled representations. The disentanglement plays a key role in interactive fashion retrieval for obtaining more controllability of the search results since it is possible to apply required operations in the desired subspace without affecting the other subspaces.

1.2 Outline of the Thesis

This thesis consists of an overview and an appendix, which includes the original articles. In Chapter 2, a brief review of previous work on multi-view depth reconstruction is presented. Chapter 3 focuses on learning correlated latent representations with Gaussian processes. Different covariance functions for encoding prior of camera poses are also presented in detail. Chapter 4 studies two different ways of learning disentangled representations. Chapter 5 discusses the limitations and future directions of the research. In Chapter 6, the conclusion of the thesis is presented.

2. Deep Geometry-based Depth Reconstruction

Depth estimation from images is a fundamental problem in computer vision, and any progress in the field can have direct impacts on applications like autonomous driving and augmented reality. Given a set of images of a scene or an object, after the camera motion is recovered, multi-view stereo (MVS) aims to assign each pixel in the images a 3D point. For 3D geometry reconstruction, the related classical methods can be divided into three categories: direct point cloud reconstructions [21, 51], volumetric reconstructions [48, 73] and depth map reconstruction [5, 78]. This chapter will focus on depth map reconstruction methods.

The key of recovering the 3D geometry is to solve the correspondence matching problem across the input images. To find the corresponding pixels, two essential elements need to be considered: potential pixel candidates in other views that can be generated efficiently and the cost metric that can measure the similarity of the given candidates. Since the camera parameters are given, the correspondence matching problem can be regarded as a 1D search along the epipolar line [26]. For the similarity metrics, there are different *photo-consistency* measurements that are common for pixel-based matching [33]. Given two images I_i and I_j and a 3D point $\mathbf{p}$ that can be seen by all two images, one can compute the photometric consistency between the projections of point $\mathbf{p}$ as:

$$C_{ij} = \rho(I_i[\Omega(\pi_i(\mathbf{p}))], I_j[\Omega(\pi_j(\mathbf{p}))]), \tag{2.1}$$

where $\rho(\mathbf{f},\mathbf{g})$ is the similarity measures between input vectors $\mathbf{f}$ and $\mathbf{g}$, $\pi_i(\mathbf{p})$ is the projection of 3D point $\mathbf{p}$ in image I_i, $\Omega(\mathbf{x})$ denotes a support domain around point $\mathbf{x}$. Different photo-consistency metrics can be defined as different choice of $\rho(\mathbf{f},\mathbf{g})$ and $\Omega(\mathbf{x})$. There are popular options like the sum of square difference (SSD) and the zero-mean normalized cross correlation (NCC). The SSD computes the L2 distance between input vector $\mathbf{f}$ and $\mathbf{g}$: $\rho_{SSD}(\mathbf{f},\mathbf{g}) = ||\mathbf{f}-\mathbf{g}||^2$. Though the SSD is time-efficient, it is not robust enough to illumination gain and bias. The NCC is a commonly-used cost metric in MVS methods. It considers the illumination gain differences in

the matching domain via normalization:

$$\rho_{NCC}(\mathbf{f},\mathbf{g}) = \frac{(\mathbf{f}-\bar{\mathbf{f}})\cdot(\mathbf{g}-\bar{\mathbf{g}})}{\sigma_{\mathbf{f}}\sigma_{\mathbf{g}}},$$

where $\bar{\mathbf{f}}$ denotes the mean of $\mathbf{f}$ and $\sigma_{\mathbf{f}}$ denotes the standard derivation of $\mathbf{f}$.

To solve the correspondence matching problem across multiple images, plane sweep algorithm [14] discretizes the depth space with fronto-parallel planes, projects images onto the planes, and measures the photometric similarity of projected pixels on the planes. The projection is equivalent to a planar homography warping from the neighbour frames to the reference frame at sampled depths. Given the intrinsic matrix $\mathbf{K}$, the relative rotation $\mathbf{R}$ and the translation $\mathbf{t}$, considering two camera models $\mathbf{P} = \mathbf{K}[\mathbf{I}|\mathbf{0}]$ and $\mathbf{P}' = \mathbf{K}'[\mathbf{R}|\mathbf{t}]$ and a plane defined as $\Pi = (\mathbf{n}^\top, d)^\top$, the homography matrix can be computed as:

$$\mathbf{H} = \mathbf{K}'\left(\mathbf{R} - \frac{\mathbf{t}\mathbf{n}^\top}{d}\right)\mathbf{K}^{-1}. \tag{2.2}$$

With the $\mathbf{H}$, the homogeneous coordinates can be transformed as $\mathbf{x}' = \mathbf{H}\mathbf{x}$. For M sampled fronto-parallel planes, the plane normal can be written as $\mathbf{n}^\top = (0,0,1)$, and $d = -d_m$ where the sampled depths $d_m \in [d_{near}, d_{far}]$ for $m = 1,\ldots,M$. Under the Lambertian surfaces assumption, if the plane is located at the real surface, then the colors of warped pixels should be similar. Using a cost metric like SSD and NCC can measure the similarity, and the depth map can be extracted from the cost volume via the simple winner-takes-all strategy or designed global optimization strategies [41].

Recent success on deep learning has triggered the interest to improve the MVS with learning-based methods. To encode the geometry information, many learning-based MVS methods rely on plane sweep algorithm to build cost volumes [35, 81, 86, 87, 36]. MVdepthNet [81] pre-computes a 3D cost volume by using the absolute difference of RGB values as the cost metric, and the cost volume is then stacked with the reference image as the input to the framework. Instead of comparing the RGB difference, MVSNet [86] extracted feature maps from images firstly and then built a 3D variance-based cost volume from the warped feature maps. DPSNet [36] concatenate the warped features to obtain a 4D volume without using a distance metric. Though the concatenated features improve performance, the 3D convolution layers make the prediction slower. After aggregating the cost volume with deep neural networks, to get the final depth map, some methods turn the prediction into a classification task and use soft-argmax operation to compute the continuous depth estimation [86, 36], while some methods regress the predictions directly [81].

2.1 Histogram-Based Depth Plane Sampling

The plane sampling plays an essential role in the plane sweep algorithm. Both the number of planes and the selection of planes affect results. Generally, with more sampled planes, the accuracy of prediction will be higher, but the time to construct the cost volume will be longer [81]. For plane selection, some methods uniformly sample planes in the depth domain [86, 91]. Give the pre-defined depth range $[d_{min}, d_{max}]$, the i-th sample of N_d planes can be selected by

$$d_i = (d_{max} - d_{min}) \frac{i}{N_d - 1} + d_{min}. \tag{2.3}$$

On the other hand, some methods show that uniformly sampling in the inverse-depth domain improves performance since it provides denser samples in areas closer to the camera [36].

$$\frac{1}{d_i} = (\frac{1}{d_{min}} - \frac{1}{d_{max}}) \frac{i}{N_d - 1} + \frac{1}{d_{max}}. \tag{2.4}$$

Most of these sampling methods rely on a pre-defined fixed depth range, while methods like DeepMVS [35] pre-run the traditional methods like COLMAP [71] to estimate the depth range firstly.

To enable the model to work for both indoor scenes and outdoor scenes well with fewer planes at fixed ranges, Publication I proposes the idea of sampling planes based on the cumulative histogram of depths, which provides enough coverage of depth planes in both nearby and far areas when the training set is a mixed dataset. The depth density and cumulative depth density function can be defined as

$$p(d_i) = \frac{n_i}{N} \quad \text{and} \quad P(d_i) = \sum_{j=1}^{i} p(d_i), \tag{2.5}$$

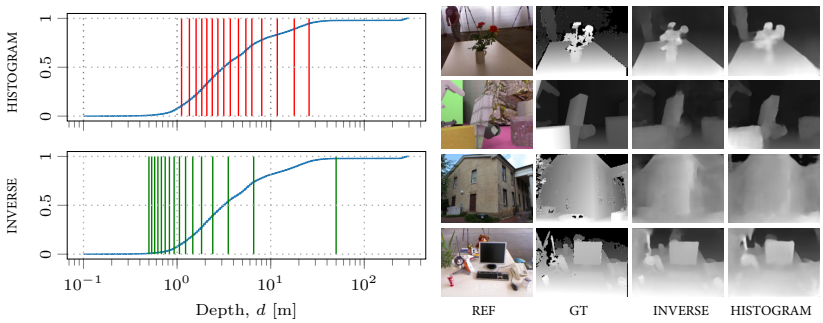
where n_i is the frequency of the depth value d_i, and N is the number of pixels with valid ground truth depth values in the training dataset. Given the cumulative density function P, a set of depth values can be sampled with the inverse density function

$$d_i = P^{-1}(\theta_i), \tag{2.6}$$

where $\theta_0, \theta_1, ..., \theta_{N_d - 1}$ are uniformly selected quantiles of P within range $[\theta_{min}, \theta_{max})$

$$\theta_i = (\theta_{max} - \theta_{min}) \frac{i}{N_d} + \theta_{min}. \tag{2.7}$$

Figure 2.1 shows the comparison between histogram-based sampling and inverse depth based sampling. The main motivation of sampling uniformly in the inverse-depth domain is to have a higher density for close depths.



(a) Two ways of depth plane sampling (b) Comparison of disparity maps

Figure 2.1. Comparison between two different plane sampling schemes. (a) The upper figure shows the selected planes from histogram-based sampling and the bottom figure shows selections from inverse depth based sampling. The x-axis denotes the depth in log scale, while the blue curves present the cumulative histogram of depths of the training set. (b) Qualitative comparison between inverse depth based sampling and histogram-based sampling on different datasets.

Generally, there will be more pixels in nearby areas, so using histogram-based sampling is naturally in accordance with the motivation. With limited number of planes ($N_d = 15$) and the fixed depth range $d_i \in [0.5, 50]$, most selected planes from inverse depth based sampling are located in extreme close regions ($d_i < 1$), and for distant region there is no enough planes. Differently, the histogram-based sampling provides coverage for both nearby and distant regions with $\theta_i \in [0.1, 1)$. Moreover, the curve in Figure 2.1 shows the cumulative histogram of depths. The higher slope of the curve means denser distribution of pixel depths. The selection results show that histogram-based sampling provides more depth planes within the depth range with a higher slope, and the density in closer areas is also higher than far areas. The qualitative predictions show that histogram-based sampling enables the model to work well with sparse planes in both small-scale and large-scale depth scenes. For the last row in the figure, extremely close objects get worse predictions with histogram-based sampling than inverse depth based sampling due to the lack of depth planes within 1 meter, which also shows that the plane selection has an important impact on the final predictions.

In summary, the histogram-based sampling attempts to improve the selection of potential candidates for correspondence matching. When reducing the number of candidates (depth planes in plane sweep algorithm), the plane sampling has a significant impact on the quality of predictions, and the histogram-based sampling can deal with both small-scale depth range like indoor scenes and larger-scale range in outdoor scenes with the same fixed planes, by utilizing the distribution of depth values in the

training set as a prior knowledge to guide the sampling.

2.2 Mask-based Multiplane Representation

Traditional plane-sweep based methods use hand-crafted cost metrics to measure photo-consistency of warped images to find an optimal plane for each pixel. For challenging regions like texture-less regions or non-lambertian surfaces, the designed cost metrics can fail and lead to noisy predictions. Some learning-based methods like DPSNet [36] use 3D convolution layers to aggregate concatenated feature maps directly without using a cost metric, but it can be more time-consuming.

Similar to the intermediate multiplane representations for image based rendering like Layered Depth Image (LDI) [74] and Multiplane Image (MPI) [93], Publication I proposes a MaskNet module to learn a mask-based multiplane representation without defining a cost metric. Since an MPI also includes a set of fronto-parallel planes, it shares similarities with the cost volume built by plane sweep. To render images from an MPI, each plane work as a RBGA layer that encodes an RGB image and an alpha map for the visibility. For depth estimation problem, the RGB information is not needed to present the geometry, so the proposed mask-based multiplane representation only attempts to roughly learn the opacity from the volume. In plane sweep algorithm, the core idea is to verify the sampled depth planes with the photo-consistency between warped pixels. However, with sparse discretized depth planes, many pixels may miss the optimal corresponding plane. The intuition of the mask-based multiplane representation is to predict whether the ray will hit a surface before the plane for each pixel. Given a reference frame and the warped images of a neighbour frame on two successive planes, if the relative position of a warped pixel flips, it hints the surface is between the two sampled planes. The intermediate representations learning can be then regarded as a binary classification task, and the binary ground truth can be extracted from the ground truth depth maps.

Given N_d selected depth planes, the input of the MaskNet consists of the reference images and warped images from all planes, which has size $3(1+N_d)\times H\times W$. Then there is an encoder-decoder architecture to predict the multiplane representation. A sigmoid function follows the prediction layer of the MaskNet, and the predicted value for each pixel on each sampled plane denotes the probability that the ray will hit the surface before the plane. The pixel-wise cross-entropy loss can be computed to train the MaskNet. After obtaining the multiplane representation from the MaskNet, another encoder–decoder module, DispNet, is used to regress the final depth maps. The learned multiplane representation is concatenated with the reference image as the input of the DispNet, which has the size

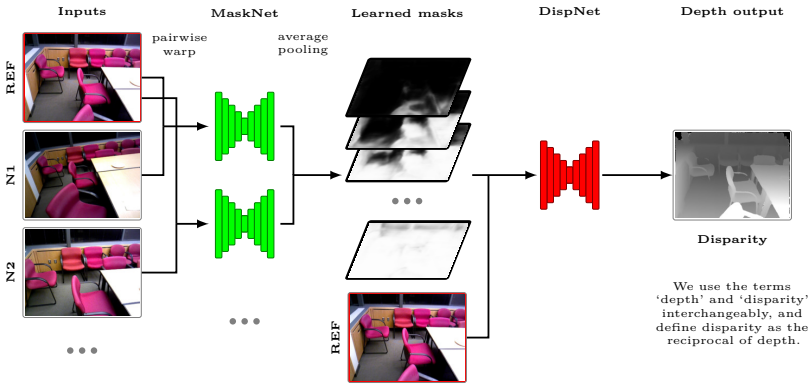


Figure 2.2. Overview of the proposed workflow in Publication I. Given the reference image and the warped neighbour images on sampled depth planes, an intermediate multiplane representation is learned firstly to predict the probability of a ray terminating before each plane. Then the continuous inverse depth maps will be predicted based on the learned masks.

$(3+N_d)\times H\times W$. The DispNet can be trained with the average L1 loss between the estimated inverse depth (disparity) and the ground truth inverse depth. Figure 2.2 shows the overview of the workflow.

Compared to the cost volumes used in other plane sweep based methods, the main difference of the proposed mask-based multiplane representation is that it does not need to compute any cost metric. Moreover, since it is not just verifying whether the sampled plane is located at the real surface but predicting the probability of ray termination, it is considering the continuous volume density with discrete samples and more suitable for sparse planes. The results reported in Table 2.1 show that when there are only 16 planes, the proposed MaskNet performs better than other state-of-the-art plane sweep based methods. As the running time will drop when using fewer planes, it can benefit real-time systems. And the significant improvement of the outdoor scenes (MVS) and synthesized scenes (scenes11) can be introduced by still considering far planes when the samples are sparse.

Table 2.1. Comparison results between MVDepthNet, DeepMVS, COLMAP, and our method. We outperform other methods in most of the data sets and error metrics (smaller better).

		MVDepth-16	DeepMVS	COLMAP	Ours (Hist.)	Ours (Inv.)
scenes11	L1-rel	0.2352	0.3755	0.7205	**0.1475**	0.2144
	L1-inv	0.0292	0.0495	0.0936	**0.0231**	0.0308
	sc-inv	0.3207	0.5810	0.7814	**0.2483**	0.2985
MVS	L1-rel	0.3835	0.8217	0.9921	0.2669	0.4030
	L1-inv	0.1384	**0.1065**	0.1812	0.1377	0.1600
	sc-inv	0.3427	0.5325	0.6892	**0.3001**	0.3100
SUN3D	L1-rel	0.1840	0.8604	1.8499	0.1797	**0.1611**
	L1-inv	0.0865	0.1317	0.4511	0.0818	**0.0808**
	sc-inv	0.2013	0.4992	1.1219	0.1916	**0.1769**
RGBD	L1-rel	0.1628	0.5066	2.2992	0.1748	**0.1572**
	L1-inv	**0.0789**	0.1717	0.5593	0.0846	0.0802
	sc-inv	0.2360	0.5238	1.2970	0.2304	**0.2062**

3. Correlated Latent Representations

Most learning-based methods make predictions per frame independently. For example, variational autoencoders (VAEs) [44] assume the data are independent and identically distributed ($i.i.d.$). In problems like multi-view depth estimation, though the cost volumes utilize neighbour frames to encode geometry information, the prediction for each frame is still independent, which may lead to inconsistency across frames easily.

On the other hand, the potential benefits of utilizing correlations between input data to obtain better predictions have received less attention, though in many tasks the inputs are correlated in time, camera poses, object identities, *etc.* For example, in applications like autonomous driving, the consecutive images in the video stream can be correlated in time as they could look very similar and they should have similar latent representations. There are methods in monocular depth estimation that utilize the temporal dependency across frames without knowing the relative camera poses [63, 82]. Sometimes, the camera poses of input images are given, and they can form the relationship between frames. Considering the standard camera projection model that is characterized by the intrinsic and extrinsic parameters

$$\begin{pmatrix} u & v & 1 \end{pmatrix}^{\mathsf{T}} \sim \mathbf{K} \begin{pmatrix} \mathbf{R} & \mathbf{t} \end{pmatrix} \begin{pmatrix} x & y & z & 1 \end{pmatrix}^{\mathsf{T}}, \tag{3.1}$$

where $(x, y, z) \in \mathbb{R}^3$ are the world coordinates. $\mathbf{R}$ is a rotation matrix and $\mathbf{t}$ is a translation vector, which make up the transformation matrix from the world coordinates to the camera coordinates. $\mathbf{K}$ is the intrinsic matrix that maps the camera coordinates to image (pixel) coordinates (u, v). According to Eq. (3.1), the scene motion and camera motion would be responsible for image appearance change over time. When given a static scene that consists of a set of fixed world coordinates with a fixed camera, the driving factor of pixel values change is the camera pose, so the images captured from close poses could have similar latent representations. Leveraging the correlations between inputs could regularize the latent space and obtain better results.

Gaussian processes (GPs) provide a probabilistic machine learning frame-

work for encoding flexible prior knowledge over functions. A Gaussian process $f(\mathbf{x})$ is a random function which can be defined in the following formalism [66]:

$$f(\mathbf{x}) \sim \mathrm{GP}(\mu(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}')), \tag{3.2}$$

where $\mu(\mathbf{x})$ is a mean function and $\kappa(\mathbf{x}, \mathbf{x}')$ is a covariance function (kernel) that can be used to define similarity and encode a correlation structure. There are some popular examples of covariance functions. For example, the Radial Basis Function (RBF) specifies the covariance between pairs of random variables

$$\kappa(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{|\mathbf{x} - \mathbf{x}'|^2}{2\ell^2}\right), \tag{3.3}$$

where the σ^2 and ℓ^2 are learnable hyperparameters.

In GPPVAE [6], the standard periodic kernel is used for 1D rotation, which warp the input angles to the unit circle $\mathbf{u}(\theta) = (\cos(\theta), \sin(\theta))$ and gives

$$\kappa(\theta, \theta') = \exp\left(-\frac{2\sin^2((\theta - \theta')/2)}{\ell^2}\right). \tag{3.4}$$

Though the standard periodic kernel encodes the nearness of 1D rotations well, there are no established functions to encode priors in SE(3) induced by 6-DoF camera poses in the standard GP toolsets. The formulation of such a covariance function can be an essential building block for applying GP priors in 3D computer vision. This chapter will explore different view covariance functions in 3D vision tasks like multi-view depth estimation, stereo matching and novel view synthesis.

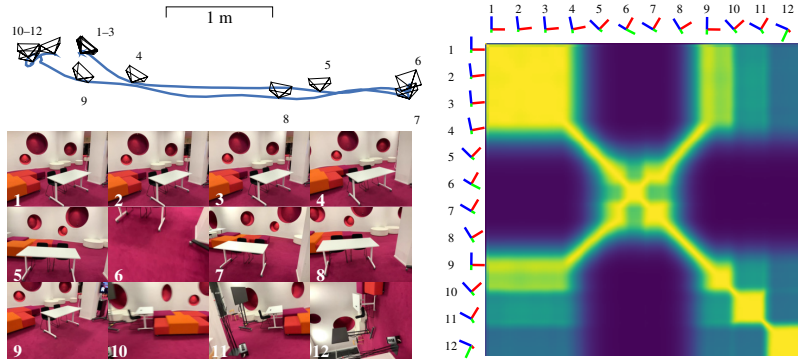
3.1 Hierarchy of View Priors

Since the camera poses can reveal the relationship between frames, it could be used to form the prior covariance functions. In many cases, full relative poses between frames can be available via odometry technology and visual-inertial systems [58, 65, 75]. But sometimes, for handheld devices with low-quality Micro Electro-Mechanical System (MEMS) only, estimating the full 6-DoF poses can be infeasible, and only 3-axis gyroscope data is available.

According to the availability of motion information, a hierarchy of three movement-induced GP prior kernels is presented in Publication III:

- when the full 6-DoF relative poses with rotation matrices and translation vectors are known, a full pose kernel can be used.

- when only angular rates of the relative orientation changes are available, a gyroscope kernel can be used.



(a) Camera pose track and frames (b) Pose-kernel in batch mode

Figure 3.1. Example of the full pose kernel proposed in Publication II. (a) A continuous camera trajectory on the left with associated camera frames. (b) The pose kernel encodes information about how much 'closeness' we expect certain views to have with their latent representations. For example, the covariance values between frames 1–4 and 9 are higher.

- even when motion is unknown, a time-decay kernel can still be used.

Full pose kernel In the mathematical sense, the camera poses belong to the special Euclidean group, SE(3), which comprise arbitrary combinations of translations and rotations $\mathrm{SO}(3) \times \mathrm{T}(3)$. The SO(3) denotes the special orthogonal rotation group, and the T(3) denotes the group of translations.

One way to build the covariance function of full camera poses is using custom distance measures. The Publication II adopt the pose distance metric proposed by Mazzotti *et al.* [56] and Matérn class [66] to build the following covariance function:

$$D[P_i, P_j] = \sqrt{\|\mathbf{t}_i - \mathbf{t}_j\|^2 + \frac{2}{3}\mathrm{tr}(\mathbf{I} - \mathbf{R}_i^\top \mathbf{R}_j)}, \tag{3.5}$$

$$\kappa(P_i, P_j) = \gamma^2 \left(1 + \frac{\sqrt{3}D[P_i, P_j]}{\ell}\right) \exp\left(-\frac{\sqrt{3}D[P_i, P_j]}{\ell}\right). \tag{3.6}$$

The $\mathrm{tr}(\cdot)$ gets the trace of the matrix, and the learnable hyperparameters γ^2 and ℓ define the characteristic magnitude and length-scale of the Gaussian processes. Fig. 3.1 shows an example of the full pose kernel computed by Eq. (3.6). Frames that look similar have closer camera poses and higher covariance values.

Publication IV proposes another novel full pose kernel that considering translation and rotation separately $\kappa(P_i, P_j) = \kappa_{trans}(\mathbf{t}_i, \mathbf{t}_j)\kappa_{rot}(\mathbf{R}_i, \mathbf{R}_j)$. As translation vectors reside in $\mathbb{R}^3$, the RBF covariance function Eq. (3.3) can be applied directly to get

$$\kappa_{trans}(\mathbf{t}_i, \mathbf{t}_j) = \sigma^2 \exp\left(-\frac{|\mathbf{t}_i - \mathbf{t}_j|^2}{2\ell^2}\right). \tag{3.7}$$

To formulate a proper covariance function in SO(3) for rotations, since the Euler angles can suffer from gimbal lock [16, 19]) and the quaternion can suffer from non-uniqueness, the covariance function of rotation matrices is proposed. Considering the eigendecomposition of $\mathbf{R}$ that define the rotation axis and angle, the geodesic distance can be defined as

$$d_{\mathrm{g}}(\mathbf{R},\mathbf{R}') = \arccos\left(\frac{1}{2}(\mathrm{tr}(\mathbf{R}^{\mathsf{T}}\mathbf{R}')-1)\right). \tag{3.8}$$

Then a Taylor expansion gives $d_{\mathrm{g}}(\mathbf{R},\mathbf{R}') \approx \sqrt{\mathrm{tr}(\mathbf{I}-\mathbf{R}^{\mathsf{T}}\mathbf{R}')}$, and the covariance function can be written as

$$\kappa_{\mathrm{view}}(\mathbf{R},\mathbf{R}') = \exp\left(-\frac{\mathrm{tr}(\mathbf{I}-\mathbf{R}^{\mathsf{T}}\mathbf{R}')}{2\ell^2}\right). \tag{3.9}$$

When there is only rotation around one axis, the covariance function Eq. (3.9) leads to the 1D standard periodic kernel Eq. (3.4).

Gyroscope kernel The covariance functions of full camera poses require the pose estimation to be solved on the fly, which can be computationally expensive for low-end devices. To relax the requirement for the full poses, based on the observations of angular velocity from a gyroscope $\boldsymbol{\omega} = (\omega_{\mathrm{x}}, \omega_{\mathrm{y}}, \omega_{\mathrm{z}})$, a gyroscope kernel is proposed in Publication III.

A vector $\mathbf{r}$ undergoing uniform circular motion around an axis satisfies $\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} = \boldsymbol{\omega} \times \mathbf{r}$, which can be equivalently expressed with the infinitesimal rotation matrix

$$[\boldsymbol{\omega}]_{\times} \overset{\mathrm{def}}{=} \begin{pmatrix} 0 & -\omega_{\mathrm{z}} & \omega_{\mathrm{y}} \\ \omega_{\mathrm{z}} & 0 & -\omega_{\mathrm{x}} \\ -\omega_{\mathrm{y}} & \omega_{\mathrm{x}} & 0 \end{pmatrix}. \tag{3.10}$$

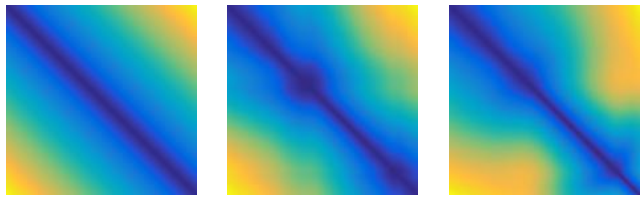
Considering the distance metric in Eq. (3.6), if an initial rotation $\mathbf{R}(0)$ is given, the rotation part can be replaced as

$$\begin{aligned} &\mathbf{R}(t)^{\mathsf{T}}\mathbf{R}(t') \\ &= \left[\int_0^t [\boldsymbol{\omega}(\tau)]_{\times}\mathbf{R}(0)\,\mathrm{d}\tau\right]^{\mathsf{T}}\left[\int_0^{t'} [\boldsymbol{\omega}(\tau)]_{\times}\mathbf{R}(0)\,\mathrm{d}\tau\right] \\ &= \int_t^{t'} [\boldsymbol{\omega}(\tau)]_{\times}\mathbf{I}\,\mathrm{d}\tau. \end{aligned} \tag{3.11}$$

Assuming a piece-wise constant rotational rate, given a collection of time-stamped angular velocity observations $\{(t_k, \boldsymbol{\omega}_k)\}_{k=i}^{j}$ between time instances t_i and t_j, the distance function for rotations can be written as

$$d_{\mathrm{gyro}}(t_i, t_j) = \sqrt{\mathrm{tr}(\mathbf{I}_3 - \textstyle\prod_{k=i+1}^{j} \exp(-[\boldsymbol{\omega}_k]_{\times}\,\Delta t_k))}, \tag{3.12}$$

where $\Delta t_k = t_k - t_{k-1}$ and 'exp' denotes the matrix exponential function. We disregard any possible additive or multiplicative biases in this distance metric, and simply assume the gyroscope to be suitably calibrated.



(a) $\mathbf{D}_{\mathrm{t}}$ (b) $\mathbf{D}_{\mathrm{gyro}}$ (c) $\mathbf{D}_{\mathrm{pose}}$

Figure 3.2. Examples of three levels of the hierarchy of view priors. The markovian gyroscope kernel captures a similar pattern as the pose distance without access to the full pose information.

Moreover, to define the rotational distance between input frames in a Markovian fashion, the cumulative distance $s_i = \sum_{j=1}^{i} d_{\mathrm{gyro}}(t_{j-1}, t_j)$ is computed firstly, and then the proposed gyroscope kernel can be written as

$$\kappa_{\mathrm{gyro}}(t_i, t_j) = \gamma^2 \left(1 + \frac{\sqrt{3}\,|s_i - s_j|}{\ell}\right) \exp\left(-\frac{\sqrt{3}\,|s_i - s_j|}{\ell}\right), \tag{3.13}$$

where γ^2 and ℓ are learnable hyperparameters.

Time-decay kernel Even there is no available motion data, since the consecutive frames captured from close timestamps are more likely to look similar and vice versa, the time difference between input frames can also be utilized to form the correlation structure. In that case, the time-decay kernel can be written as a stationary Matérn covariance function

$$\kappa_{\mathrm{t}}(t, t') = \gamma^2 \left(1 + \frac{\sqrt{3}\,|t - t'|)}{\ell}\right) \exp\left(-\frac{\sqrt{3}\,|t - t'|)}{\ell}\right), \tag{3.14}$$

where γ^2 and ℓ are learnable hyperparameters, and t and t' denote timestamps. Figure 3.2 shows the examples of different kernels based on different levels of motion information.

3.2 Latent Nonparametric Fusion

Autoencoder architectures play crucial roles in learning low-dimensional latent representations [32]. Different from principal components analysis (PCA) that only considers linear transformations, autoencoders use a nonlinear multilayer encoder network to transform the high-dimensional data (*e.g.*, images) into a low-dimensional code. Then a decoder network is used to predict output results from the latent code. Since there is information lost during downsampling, to refine fine-grained details better in outputs, the skip connections between encoder layers and decoder layers are used to pass low-level features [69]. Combined with convolutional neural networks (CNNs), the encoder–decoder architectures are widely

used for diverse computer vision problems like image segmentation [69, 4], depth estimation [22, 81], image synthesis [44, 9], *etc.*

To take the correlation between inputs into consideration, the GP priors can be introduced in the latent code. Considering the extracted latent representation $\mathbf{y}_i$ from the encoder as noise-corrupted observation of the 'ideal' latent representations $\mathbf{z}_i$, given the view prior based on the motion information, the 'ideal' latent representations $\mathbf{z}_i$ can be inferred via the GP regression model

$$\begin{aligned} \mathbf{z}_{i,j} &\sim \mathrm{GP}(0, \kappa(P_i, P_{i'})), \\ \mathbf{y}_{i,j} &= \mathbf{z}_{i,j} + \varepsilon_{i,j}, \quad \varepsilon_{i,j} \sim \mathrm{N}(0, \sigma^2), \end{aligned} \tag{3.15}$$

where i is the index of the input frame, P_i denotes the camera pose data, and j is the index of the dimension of latent representations. The noise variance σ^2 is a learnable parameter. The GP priors are assigned independently to each dimension of the latent representations.

To obtain the 'ideal' latent representations, the posterior mean can be computed by solving the GP regression with one matrix inversion [66]:

$$\begin{aligned} \mathbb{E}[\mathbf{Z} \mid \{(P_i, \mathbf{y}_i)\}_{i=1}^N] &= \mathbf{C}(\mathbf{C} + \sigma^2 \mathbf{I})^{-1} \mathbf{Y}, \\ \mathbb{V}[\mathbf{Z} \mid \{(P_i, \mathbf{y}_i)\}_{i=1}^N] &= \mathrm{diag}(\mathbf{C} - \mathbf{C}(\mathbf{C} + \sigma^2 \mathbf{I})^{-1} \mathbf{C}), \end{aligned} \tag{3.16}$$

where $\mathbf{Z} = (\mathbf{z}_1 \ \mathbf{z}_2 \ \dots \ \mathbf{z}_N)^\top$ are stacked 'ideal' latent representations, $\mathbf{Y} = (\mathbf{y}_1 \ \mathbf{y}_2 \ \dots \ \mathbf{y}_N)^\top$ are stacked extracted latent representations from the encoder, and $\mathbf{C}$ denotes the prior covariance matrix $\mathbf{C}_{i,j} = \kappa(P_i, P_j)$. Then instead of the extracted latent representations from the encoder, the posterior mean for each frame $\mathbb{E}[\mathbf{z}_i \mid \{(P_i, \mathbf{y}_i)\}_{i=1}^N]$ is passed to the decoder to get the output result. Fig. 3.3 shows an example sketch of the latent nonparametric fusion with the encoder-decoder architecture. The parameters of the encoder network, the decoder network and the hyperparameters of GP priors are optimized jointly during training.

The GP regression in the latent space allows soft fusion between multiple frames by leveraging the correlation between input data. In Publication II, the experimental results show that the latent nonparametric fusion leads to more temporal consistent predictions for multi-view depth estimation. Moreover, it is more robust to ill-conditioned inputs. For example, for multi-view depth estimation problem, when the baseline between two images is too small, the cost volume cannot encode geometry information. In that case, the vanilla encoder-decoder architectures like MVDepthNet [81] that take the cost volume as a part of input may fail, while our predictions can be more robust as the neighbour latent representations can be propagated to 'recover' the failed latent representation. Similar improved results also are observed for stereo matching problem with the gyroscope kernel in Publication III.

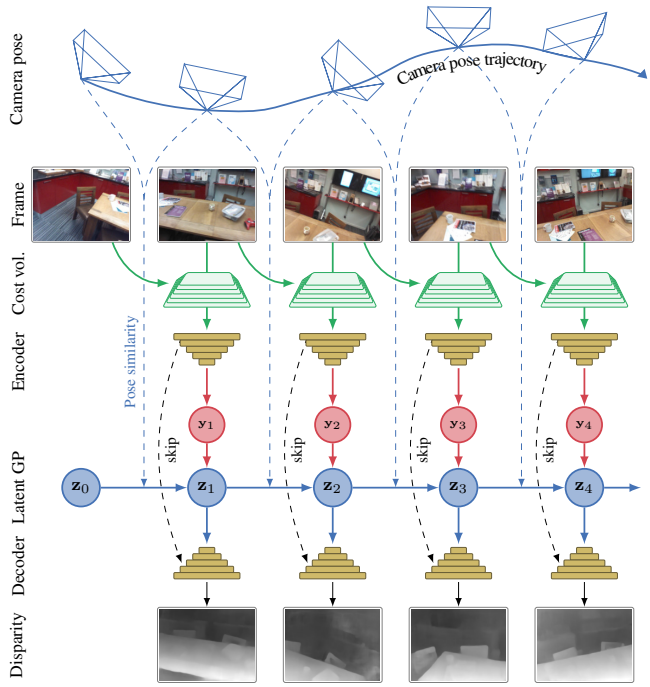


Figure 3.3. The workflow of GPMVS that uses latent nonparametric fusion to solve the MVS problem better. The current frame and the previous frame are used to compute a cost volume, which is then stacked with the current frame as the input to the encoder network. After extracting latent representations from the encoder, the pose priors are introduced and the latent representations are updated via the GP inference. The new latent representations will be passed through the decoder network to get the final outputs. Figure adapted from Publication II.

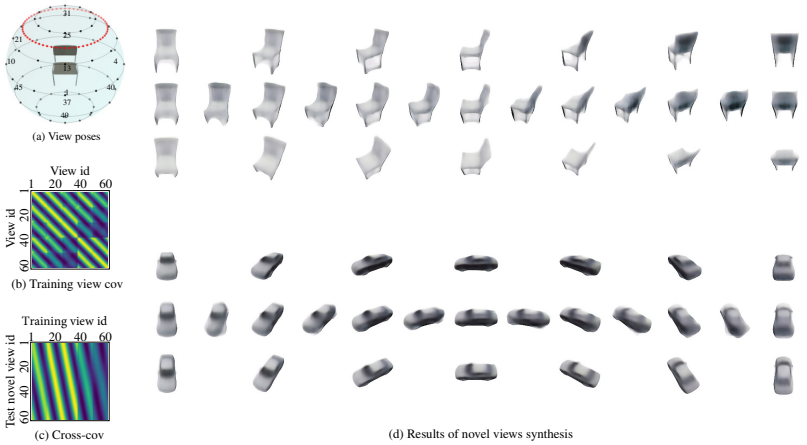


Figure 3.4. Experiment results on ShapeNet with GPPVAE and the proposed view kernel. (a) Black dots denote the 60 view angles in training data, while red dots denote the novel test views. (b) The covariance matrix for 60 training views. (c) The cross covariance matrix for the training views and test views. (d) Generated images for a chair and a car. For each category, the first (elevation 30°) and the third row (elevation 60°) show predictions for views from angles found in the training set, while the second row shows synthesis results for angles never observed for any objects in the training set.

The latent nonparametric fusion can also be applied to generative models to synthesize images. Instead of using the iid prior over the latent space, GPPVAE [6] considers the GP prior that is factorized into an object kernel and a view kernel. Though GPPVAE extends the regular VAEs with the power of modeling correlations in latent space, their view kernel only supports 1D arbitrary angles or fixed angles. By only replacing the view kernel with the proposed view kernel that supports 3D rotations in SO(3) in Eq. (3.9), Publication IV enables the GPPVAE to synthesize novel views with arbitrary 3D angles. The resulting composite kernel can be written as

$$\kappa(\mathbf{x},\mathbf{R};\mathbf{x}',\mathbf{R}') = \underbrace{\mathbf{x}^\top\mathbf{x}'}_{\text{object}}\underbrace{\exp\Big(-\frac{1}{2}\mathrm{tr}(\mathbf{\Lambda}-\mathbf{R}^\top\mathbf{\Lambda}\mathbf{R}')\Big)}_{\text{view}}, \tag{3.17}$$

where $\mathbf{\Lambda} = \mathrm{diag}(\ell_\mathrm{x}^{-2},\ell_\mathrm{y}^{-2},\ell_\mathrm{z}^{-2})$ and $\ell_\mathrm{x},\ell_\mathrm{y},\ell_\mathrm{z}$ are learnable lengthscale hyperparameters. $\mathbf{x}$ and $\mathbf{x}'$ are learned object feature vectors, and $\mathbf{R}$ and $\mathbf{R}'$ are rotation matrices. Fig. 3.4 shows qualitative experiment results on ShapeNet [7] datasets. Compared to frame-independent inference, leveraging the closeness of camera poses as the GP priors regularizes the latent representation softly and makes the latent space more interpretable.

4. Disentangled Representation Learning

Though deep learning shows promising potential in learning low-dimensional latent representations, the lack of the controllability and interpretability may limit the application in real-world problems. A solution is to disentangle the underlying *factors of variation* [29, 55, 68, 95]. For instance, for generating an image of an object, computer graphics models suggest controlling factors like object identity, object location, pose, lighting, *etc.* Ideally, the factors of variation only change some specific characteristics of the data, while leaving all other characteristics invariant. Based on the invariance, the disentangled representations can benefit some tasks that require the representations to be unaffected by uninformative factors, such as pose-invariant face recognition [64, 79]. In the context of generative models for image synthesis, the disentangled representations provide flexible control of synthesis results [59, 28, 42, 43, 80, 45]. For example, StyleGAN [42] can separate the fine-grained attributes (*e.g.*, hair, freckles) from higher-level variation (*e.g.*, identity, pose). Different from the coarse-to-fine level control, some methods provide explicit control over specific factors. HoloGAN [59] can control the pose of generated object images via disentangling 3D pose and identity. ConfigNet [45] can generate face images with desired attributes such as head pose, smile, hair color, presence or absence of eyeglasses, *etc.*

To learn disentangled representations, many methods decompose latent representations into independent subspaces with respect to factors of variation [34, 77, 47]. When there is prior knowledge about the data domain, and the factors of variation are labelled in the training set, the disentangled representation learning can be then supervised by the labels [12, 88, 90, 67]. For instance, as it is common to disentangle the pose and identity in 3D vision [61, 77, 85]. Tatarchenko *et al.* concatenate the extracted latent representations from the input image and the target angles to generate novel view images [77]. Peng *et al.* propose a feature reconstruction metric learning method to disentangle identity and pose for better face recognition performance [64]. The Publication VI leverages the semantic feature labels of clothing images to learn attribute-specific

representations to improve the interactive retrieval results.

Besides factorizing latent representation into independent chunks, equivariant representations are also able to disentangle the factors of variation. Different from many methods based on factorisation that attempt to make sub-features invariant to other transformations (*e.g.*, pose change), the equivariant representations aim to preserve information about the transformation [84, 31, 60, 50, 59, 13]. Hinton *et al.* proposed transforming auto-encoders (TAE) to model both 2D and 3D transformations, where the transformation matrix can be applied to the output features of the capsules directly [31]. Daniel *et al.* introduce the feature transform layer to simulate many image-space transformations [84]. Transformable bottleneck network [60] also applies spatial transformations to a volumetric bottleneck directly. Based on TAE, Chen *et al.* propose an image-based rendering method that provides continuous view control for novel view synthesis [9]. Inspired by [9], Publication V learns a transformable latent representation to regress pixels of novel views with depth-guided skip connections.

Considering the challenges of obtaining labels for factors of variation, some methods attempt to learn disentangled representation in an unsupervised manner [34, 30, 8]. A significant fraction of unsupervised methods are based on generative modelling. The VAE based methods like β-VAE use an extra penalty on the KL-divergence term to match the Gaussian prior. InfoGAN [8] learn interpretable latent representations by maximizing the mutual information between synthesized images and latent features. Some methods swap out features explicitly and propose invariance loss functions to encourage disentanglement [28, 34, 39]. For example, Hu *et al.* learn disentangled feature chunks by mixing and unmixing encoded features with autoencoders [34]. Though Publication VI relies on the supervisory signal of fashion attributes, it also leverages the idea of feature mixing during the training to maintain disentanglement.

4.1 Equivariant Representation for Novel View Synthesis

Given single or multiple input images of a scene, the novel view synthesis (NVS) is the task that aims to generate new images of the scene from a novel viewpoint. Generally, except for the 3D model-based rendering techniques, previous methods for NVS can be divided into two main classes, image-based rendering (IBR) or pixel generation methods. Image-based rendering methods use a collection of given images to render novel views rather than geometric primitives, which can re-use pixels from source images directly. Some IBR systems do not require explicit geometry but use the correspondences between images. For example, learning-based methods like [40, 94, 76] predict the appearance flow directly. When the depth

maps are available, the image can be rendered from nearby viewpoints via 3D warping [54], which elevates pixels to 3D points and re-projects them onto new images to get pixel-to-pixel correspondences. Some IBR methods also leverage deep learning to predict depths from images [20, 9] when there are no depths given as inputs. Since pixels in the input source views can be re-projected to a novel view, the IBR methods can preserve original low-level details of the scene such as textures and colors well. However, obtaining accurate correspondences between images is challenging. When the source input is just a single image, the depth prediction can be even harder. And the failure can easily lead to distortion in generated new images. On the other hand, some learning-based methods attempt to predict pixel colors in target views instead of re-using the pixels [77, 85]. Since the goal is to synthesize the appearance of the scene seen from other viewpoints, it is natural to disentangle the camera poses and the appearance. With the disentangled representations of images, these pixel generation methods can generate structurally consistent geometric shapes, but the visual quality of the generated images is worse than IBR methods due to the lack of low-level details.

To combine the advantages of both IBR methods and pixel generation methods, Publication V proposes a pixel generation method that uses warped feature maps with skip connection layers to preserve low-level details. Different from other pixel generation methods that extract latent representations for pose and appearance separately, we adopt equivariant representation as [9] to explicitly control the viewpoint of the generated results.

4.1.1 Transformable Latent Representations

Chen *et al.* proposed an IBR method that learns an equivariant representations to predict depth maps for target views directly [9]. Given the source image input I_s, the encoder network ϕ extracts the latent code $z_\mathrm{s} = \phi(I_\mathrm{s})$. Similar as TAE, the latent code can be formed as $z_\mathrm{s} \in \mathbb{R}^{n\times 3}$, and the given transformation between source viewpoint and the target viewpoint $T_{\mathrm{s}\to\mathrm{t}} = [R\,|\,t]_{\mathrm{s}\to\mathrm{t}}$ can be applied to the latent code with matrix multiplication to get the transformed latent code for the target view:

$$\tilde{z}_\mathrm{t} = T_{\mathrm{s}\to\mathrm{t}}\dot{z}_\mathrm{s}, \tag{4.1}$$

where $\dot{z}_\mathrm{s}$ is the homogeneous representation of z_s. However, monocular depth estimation with the equivariant representations is challenging and unstable predictions lead to distorted target images.

Different from [9] that only use the equivariant representations to predict depth maps for target views, as a pixel generation method, Publication V aims to predict target images from the equivariant representations with a decoder network. Considering the equivariant representation z_s which

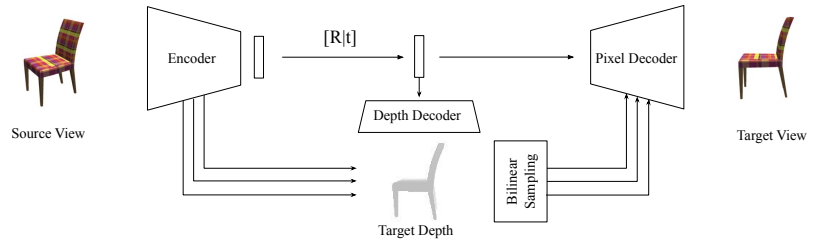


Figure 4.1. Overview of the proposed architecture in Publication V. The encoder extracts the equivariant representation, which can be multiplied with the relative transformation matrix to get the representation of the target novel view. The depth decoder and the pixel decoder will map the transformed representation to the target depth map and the target view image correspondingly.

can be regarded as a set of sparse points in latent space, intuitively the framework will encourage the latent code to encode 3D shape information implicitly. Though the sparse latent code that encodes 3D position predictions for features might be sufficient to estimate depth maps, generating novel view images with rich textures or many small objects is still challenging due to the information loss.

4.1.2 Depth-guided Skip Connections

To alleviate the lack of fine-grained details in outputs, many methods leverage skip connections layers between the encoder and the decoder to transfer low-level feature maps [69]. However, unlike tasks like semantic segmentation where the outputs and the input images are well-aligned spatially, the output images have very different shapes than the input source images. In that case, the skip connections cannot be applied immediately. As discussed in the previous section, the equivariant representations can predict depth maps for novel target views, which can be used for 3D warping. Instead of warping image pixels directly, Publication V warps the multi-level feature maps before passing them to the pixel generation decoder. As shown in Figure 4.1 , there are two decoders in the framework. The depth decoder ψ_d maps the transformed representation to the depth map $\tilde{D}_\mathrm{t} = \psi_\mathrm{d}(\tilde{z})$, and the pixel decoder ψ_p predicts the RGB values for pixels given the transformed representation and the warped multi-level feature maps. To warp the feature maps, given the camera intrinsic matrix K, the relative transformation matrix $T_{\mathrm{t}\to\mathrm{s}}$, and the predicted depth map $\tilde{D}_\mathrm{t}$, for a pixel in the target view p_s, the correspondence in the source view can be found by

$$p_\mathrm{s} \sim K\, T_{\mathrm{t}\to\mathrm{s}}\, \tilde{D}_\mathrm{t}(p_\mathrm{t}) K^{-1} p_\mathrm{t}, \tag{4.2}$$

where p_t and p_s denote the homogeneous coordinates. The differentiable bilinear interpolation [37] is applied since the obtained p_s are continuous values. The depth decoder is trained in an unsupervised manner so there

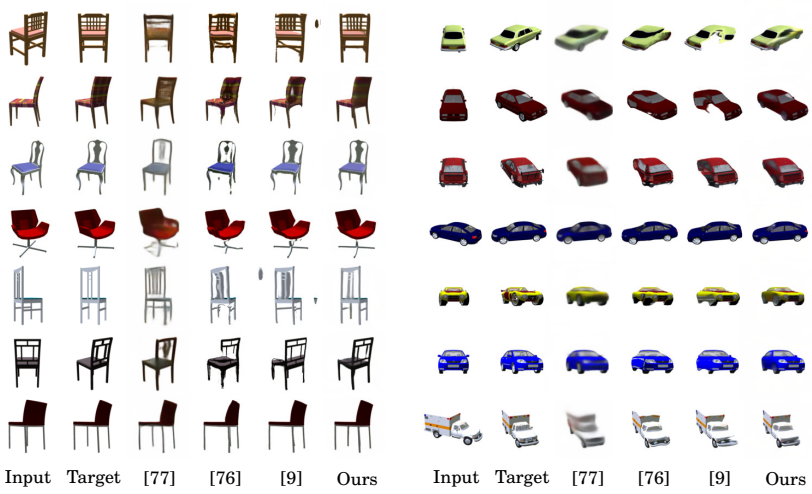


Figure 4.2. Results on different ShapeNet objects. For each object, the first two columns correspond to the input source image and the target image. The 3rd column shows the result from the pixel generation method [77]. The 5th column corresponds to the result from the IBR method [9]. The final column shows the result from the proposed method in Publication V.

is no requirement for ground truth 3D information.

The depth-guided skip connections enable the method to benefit from establishing explicit correspondences to maintain fine-grained details as IBR methods. On the other hand, as pixel generation methods, it exploits implicit learned prior to generate structure-consistent predictions to 'correct' distortion. The qualitative comparison on ShapeNet [7] objects shown in Figure 4.2 demonstrates that the proposed framework can combine the advantages of IBR methods and image generation methods. The pixel generation method [77] generates blurry results. Also, because of the lack of low-level details, the identity of the object fails to be preserved. The synthesis results from the IBR method [9] suffer from distortion. Differently, the proposed method generates structure-consistent predictions as pixel generation methods (*e.g.*, it can generate the missing chair legs and missing tires), and the generated results also include rich fine-grained details as the IBR method.

Publication V explores the use of the equivariant representation for novel view synthesis. As the pixel changes between rendered images for the given scene are driven by the viewpoint changes, it is reasonable to build the relationship between viewpoints and the latent representation in a similar way to disentangle pose and appearance. Moreover, considering the limitation of the compact equivariant representation, Publication V also exploits the benefits of explicit correspondences in feature maps to maintain fine-grained details in predictions.

4.2 Attribute-driven Disentangled Representation for Image Retrieval

When there is prior knowledge about the data domain, it is easier to disentangle the factors of variation. For example, in the fashion domain, there are diverse semantic visual attributes (*e.g.*, category, color, length of sleeves) to describe fashion items, which can be used to learn disentangled representations. Publication VI shows that the attribute-driven disentangled representations can benefit interactive fashion retrieval tasks. Interactive image retrieval for online fashion shopping provides the ability to change retrieval results according to user feedback [10, 24, 89]. It is not only necessary to obtain expressive image representations, but also to empower the model to modify the representation flexibly to meet the requirements from users. For interactions that only involve a specific aspect of the image, image representations that are semantically entangled can cause other aspects to change inadvertently (*e.g.*, a user wants to change the color of a T-shirt only, but the new retrieved results also have different sleeve type). Since the entangled representations limit the controllability of the retrieval, Publication VI uses semantic attribute labels to train deep neural networks to learn attribute-specific representations.

4.2.1 Attribute-specific Subspaces

To create a model that disentangles semantic factors in independent subspaces, the visual attributes can be used as supervisory signals. Assuming there is a predefined list of attributes of length A (*e.g.*, color, category, fabric) indexed with the symbol a. Each attribute a has a predefined list of attribute values $(v_a^1, v_a^2, ..., v_a^{J_a})$, where J_a denotes the number of possible values. Deep CNNs like AlexNet [46] or ResNet [27] can be used as the backbone network to extract the global representation $\mathbf{f}_n$ of the input image I_n firstly. To decompose the $\mathbf{f}_n$ into attribute-specific subspaces, for each attribute a there is a fully-connected two-layer network ϕ_a that maps $\mathbf{f}_n$ to a attribute-specific representation $\mathbf{r}_{n,a} = \phi_a(\mathbf{f}_n)$. The proposed architecture that extracts the disentangled representation is called the Attribute-Driven Disentangled Encoder (ADDE).

To utilize the semantic attribute labels, there is a classification layer consists of a fully-connected layer with softmax function to predict the attribute the values $\hat{y}_{n,a} = softmax(FC(\mathbf{r}_{n,a}))$. Then the training of the attribute-specific subspaces can be supervised by the cross-entropy loss for independent multi-label attribute classification tasks in subspaces as

$$L_{cls} = -\sum_{n=1}^{N}\sum_{a=1}^{A}\log(p(y_{n,a}|\hat{y}_{n,a})), \tag{4.3}$$

where $y_{n,a}$ is the ground truth label of the image I_n for attribute a, $\hat{y}_{n,a}$

is the output from the classification layer, and N denotes the number of input samples.

After extracting attribute-specific representations $\mathbf{r}_{n,a}$ for each attribute a, the final disentangled representation of the input image I_n is obtained by simply concatenating the attribute-specific representations together $\mathbf{r}_n = (\mathbf{r}_{n,1}, \ldots, \mathbf{r}_{n,A})$, where $\mathbf{r}_n \in \mathbb{R}^{A \cdot d}$ and d denotes the dimension of each attribute-specific representations. Since the dimensions separate different attributes, it is flexible to modify the final representation $\mathbf{r}_n$ when a user wants to change specific attributes of retrieval results by applying operators directly on the desired subspace without affecting the other subspaces.

4.2.2 Block-diagonal Memory Block

After training with the classification task, the extracted disentangled representations can be used as discriminative representations to retrieve similar images that share the same attribute values as the query image. However, to interact with user feedback like changing specific attributes, the model have to learn how to modify the relevant dimensions.

Attribute manipulation retrieval in the fashion domain is a new research problem. In many scenarios, given the query image, users may want to manipulate the query image interactively (*e.g.*, change the color of the dress from pink to blue) rather than simply searching for similar images as the query image. AMNet [89] proposes a memory block module that consists of a memory and a neural controller, which is followed by a manipulation fusion module. FashionSearchNet [1] utilizes attribute locations to learn attribute representations separately. More recent methods leverage generative models like GANs to synthesize images that meet attribute manipulation requirements [3, 49, 52], but for image retrieval the image generation is not necessary. Most of existing methods do not consider disentanglement. For instance, AMNet uses fully-connected layers to fuse the retrieved representation from memory block and the original image representation to obtain the final representation. FashionSearchNet also learns the global representations that fuse different subspaces with additional functions, which introduces entanglement.

To formulate attribute manipulation retrieval, for each image, it has associated attribute values that can be written as a merged single list $\mathbf{v} = (v^1, v^2, \ldots, v^J)$, where $J = \sum_{a=1}^{A} J_a$. The corresponding attribute values v^j can be present as a one-hot encoding: attribute values present in the image are encoded as 1s, the rest with 0s. Then given a query image I_q with attribute values $\mathbf{v}_q = (v_q^1, v_q^2, \ldots, v_q^J)$, the task is to find target images that have attribute description as $\mathbf{v}_p = (v_p^1, v_p^2, \ldots, v_p^J)$, which differs from $\mathbf{v}_q$ only for a subset of attributes.

Inspired by AMNet, Publication VI introduces a memory block module to store prototype representations for each attribute value. For example, for

the color attribute, there is a prototype representation for each pre-defined color in the dataset. The memory block can be initialized by averaging the attribute-specific representations of those training images with the same value. Due to the disentanglement, only relevant dimensions in the subspaces are selected to form the prototype representations. And the resulted initial memory block is a block-diagonal matrix as

$$\mathcal{M} = \begin{pmatrix} \mathbf{e}_1^1 & \dots & \mathbf{e}_1^{J_1} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{e}_2^1 & \dots & \mathbf{e}_2^{J_2} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{e}_A^1 & \dots & \mathbf{e}_A^{J_A} \end{pmatrix}, \tag{4.4}$$

where $\mathbf{e}_a^j$ stored in the columns of the memory block is the prototype representation for the j-th attribute value of the attribute a.

With the memory block module, the attribute-specific subspaces can be modified freely. Figure 4.3 shows the proposed framework for attribute manipulation. For each query with a generated manipulation requirement, there is a positive sample (target image with desired attribute values) and a negative sample. For the query image and the target image, the disentangled representations $\mathbf{r}_q$ and $\mathbf{r}_p$ are extracted via the ADDE respectively. The manipulation requirement can be encoded as a vector $\mathbf{i} = \mathbf{v}_p - \mathbf{v} - q = (i^1, i^2, \dots, i^J)$, where $i \in \{-1, 1, 0\}$ denotes removing the attribute value, adding the attribute value or keeping the value unchanged. Then the target modified representation $\mathbf{r}'$ for retrieval can be computed as

$$\mathbf{r}' = \mathbf{r}_q + \mathcal{M}\mathbf{i}. \tag{4.5}$$

As the retrieval is to find the top-K nearest neighbour based on the extracted disentangled representations, the triplet loss [72] is adopted to regularize the distance between representations. The compositional triplet loss will encourage the target modified representation $\mathbf{r}'$ to be close to the image representation extracted from the positive sample that has desired attributes:

$$L_{ct} = \max(0, d(\mathbf{r}', \mathbf{r}_p) - d(\mathbf{r}', \mathbf{r}_n) + m), \tag{4.6}$$

where $\mathbf{r}_p, \mathbf{r}_n$ are the extracted disentangled representations of the positive sample and the negative sample, m is the margin parameter, and $d(\cdot)$ is the L2 distance.

During the training, the ADDE and the memory block module will be optimized jointly. To maintain the disentanglement when updating the memory block, the block-diagonal structure needs to be preserved. In that case, a regularization loss function inspired by [70] is introduced on the non-diagonal elements:

$$L_{mem} = ||\mathcal{M} \circ \mathcal{N}||_1, \tag{4.7}$$

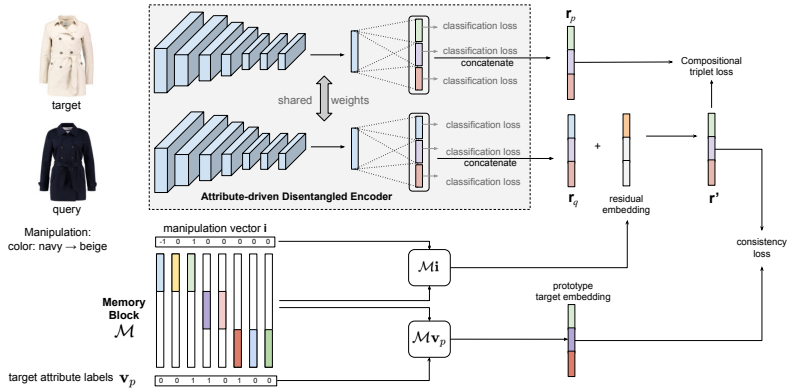


Figure 4.3. The framework proposed in Publication VI for attribute manipulation retrieval. The Attribute-Driven Disentangled Encoder (ADDE) extracts the disentangled image representations. The memory block $\mathcal{M}$ stores the prototype representations for all attribute values. The target compositional representation $\mathbf{r}'$ can be obtained by adding the residual embedding, which is computed by combining the manipulation vector $\mathbf{i}$ and the memory block.

where $\circ$ means the element-wise multiplication and $\mathcal{N}$ denotes to the non-diagonal elements as

$$\mathcal{N} = \mathbf{1}_D - \mathcal{D}, \qquad \mathcal{D} = \begin{pmatrix} \mathbf{1}_{\mathcal{M}_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{\mathcal{M}_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1}_{\mathcal{M}_A} \end{pmatrix}. \tag{4.8}$$

The $\mathbf{1}_{\mathcal{M}_a}$ is a matrix of ones of size $d \times J_a$. The L1-norm regularization loss helps to curb the mixing of different attribute-specific representations.

Moreover, due to the novel block-diagonal structure, the memory block can not only multiply with the manipulation vector to obtain residual representations for modification, but also project the one-hot attribute label vector into the disentangled subspaces directly. Intuitively, since the attribute label vector and the RGB image describe the same fashion item, the encoded semantic information should be similar, and the representation extracted from the image should be close to the prototype representation projected from the attribute label vector. To encourage this semantic consistency, there is novel loss function proposed as

$$L_c = d(\mathbf{r}_q, \mathcal{M}\mathbf{v}_q) + d(\mathbf{r}', \mathcal{M}\mathbf{v}_p) + d(\mathbf{r}', \mathcal{M}\mathbf{v}_n), \tag{4.9}$$

where $\mathbf{v}_q$, $\mathbf{v}_p$, $\mathbf{v}_n$ are the attribute value vectors of the query image, the positive sample and the negative sample generated for the manipulation task. This semantic consistency loss encourages the alignment between the prototype representations in the memory block and the extracted image representations, so the composition will be more smooth for manipulation.

	Shopping100k		DeepFashion	
	AMNet [89]	ADDE-M	AMNet [89]	ADDE-M
NDCG@30	0.7148	**0.7367**	0.2821	**0.3291**
$NDCG_t$@30	0.4010	**0.4305**	0.3347	**0.3470**
$NDCG_o$@30	0.7571	**0.7779**	0.2947	**0.3629**
Top-10	0.2562	**0.4117**	0.1411	**0.2360**
Top-30	0.4294	**0.5981**	0.2294	**0.3152**
Top-50	0.5164	**0.6729**	0.2758	**0.3591**

Table 4.1. Quantitative results on Shopping100k [2] and DeepFashion [53] datasets for attribute manipulation retrieval. Table adapted from Publication V.

Also, the prototype representation can be regarded as pseudo-supervision for the attribute-specific representation learning, which can speed up the convergence during training.

To better measure the ability to preserve attributes that should not be modified, besides using top-k retrieval accuracy as evaluation metrics, Publication VI also uses Normalized Discounted Cumulative Gain (NDCG@k) [38] and its novel variants for evaluation. The standard NDCG metric is defined as:

$$\frac{1}{Z}\sum_{j=1}^{k}\frac{2^{rel(j)-1}}{\log(j+1)}, \tag{4.10}$$

where $rel(j)$ is the number of matching attributes between the ground-truth label of j-th ranked image and the desired label divided by the total number of attribute types. Z is a normalization constant. The two variants proposed in Publication VI is called $NDCG_t$ and $NDCG_o$, which have similar formula as Eq. (4.10), though they have different way of computing $rel(j)$. The $NDCG_t$ focuses on the target attribute to be modified, so the $rel(j)$ will be binary, while $NDCG_o$ only considers the rest unchanged attributes.

Publication VI leverages the given semantic labels as the prior knowledge in the data domain and explores the attribute-driven disentangled representations for interactive fashion retrieval. The introduced attribute-specific subspaces make representation manipulation more flexible, and the tailored block-diagonal memory block module with the proposed novel loss functions enables the preservation of disentanglement. Table 4.1 shows quantitative results on Shopping100k [2] and DeepFashion [53] dataset. With the same backbone architecture, compared to the AMNet that utilizes a memory block for entangled image representations, using disentangled image representations introduces a significant improvement in terms of both NDCG metrics and top-k retrieval accuracies. The better top-k accuracies and $NDCG_t$@30 shows that the disentangled representations can retrieve images with desired attributes successfully, and the

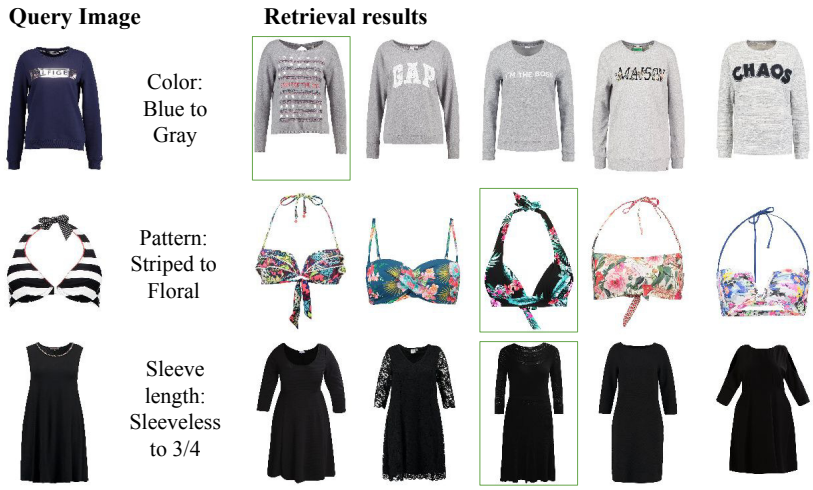


Figure 4.4. Top-5 retrieval results for attribute manipulation retrieval. The green boxes denote images that match all desired attributes.

better $NDCG_o@30$ results also indicate better preservation of undesired attributes. Figure 4.4 shows qualitative examples for attribute manipulation retrieval.

5. Discussion

This chapter summarises the findings briefly and discusses the limitations and future directions of each publication.

Chapter 2 shows a learning-based method for multi-view depth estimation. To introduce the well-studied geometric relationships as the prior knowledge to the deep architectures, a traditional plane sweep algorithm is used to build the discrete volume as a part of the input. Different from classical plane-sweep methods that use hand-crafted cost metrics to measure photo-consistency, the learned multiplane representations are used to roughly encode the opacity. Moreover, as the plane sampling affects the results of plane sweep algorithms, Publication I proposes the idea of sampling planes based on the cumulative histogram of depth. For depth ranges that have a denser distribution of pixels, the histogram-based sampling provides better coverage compared to uniform sampling. Similarly, there are other methods that also try to boost performance by improving the candidate selection for correspondence matching. For example, in NerfingMVS [83], the sparse depth maps are used to guide the point sampling for each ray. UCS-Net [11] proposes uncertainty-aware adaptive thin volumes to achieve reasonable spatial partitioning. However, since the plane sampling in Publication I will be driven by the histogram of the training dataset, the method requires the training dataset to be diverse and balanced enough. One potential direction to improve the method is automatically adjusting depth planes according to inputs. For instance, NeRF [57] utilizes hierarchical volume sampling where the second set of points are sampled based on the coarse volume distribution estimation. In Publication I, the learned multiplane representations can also estimate depth distribution roughly, which provides the possibility to vary sampled depth planes. Another limitation of Publication I is the requirement for ground truth depth labels. To further explore geometry information, some unsupervised learning frameworks for monocular depth estimation [92, 23] that use warping-based view synthesis as supervision signals can be adapted.

Chapter 3 advocates inter-frame reasoning with deep neural networks

based on the relationship between the images for image sequences instead of frame-independent predictions to obtain more consistent results. For some computer vision applications, the input images from the video stream are strongly correlated, so the correlation should be captured for latent image representations. Moreover, when camera poses of input images are given (*e.g.*, multi view stereo and novel view synthesis), the camera pose is the main factor that leads to pixel changes, so the relationship between camera poses can be utilized as the prior for latent representations. Publication II, Publication III and Publication IV build the prior covariance functions based on custom pose distance measures and leverage Gaussian process regression in the latent space to fuse multiple extracted latent image representations. Both Publication II and Publication III demonstrate that the correlated latent representations make more temporally consistent depth predictions. And Publication IV shows that using GP prior enable generative models to learn well-structured representations. Recently, more other works also realize the importance of temporal fusion. Many of them exploit recurrent networks [63, 18]. In [63], the ConvLSTM is proposed to model the spatiotemporal dependencies. Compared to the latent fusion by LSTM, the nonparametric fusion based on the Gaussian process does not need extra parameters and utilize the relationship of camera poses rather than learning the transitional kernel from scratch. The weakness of using the GP prior is that though it considers the pose changes, it only introduces soft constraints in the latent space without using any perspective correction. In that case, the errors can also be propagated during the fusion. Inspired by both Publication II and [63], DeepVideoMVS [18] propose a novel hidden state propagation scheme for ConvLSTM by explicitly warping based on perspective projection. Though the latest methods considers the temporal dependency across frames, the dynamic objects remain a open problems. Another direction of exploring the correlation of data is metric learning [25]. Loss functions like contrastive loss [25] and triplet loss [72] are designed to map similar input data to nearby points on the manifold. However, using the camera poses for metric learning received less attention. Some retrieval-based relocalization method can be improved by making use of the relationship between camera poses. For example, CamNet [17] considers the camera frustum distance when training the global descriptors, and combining the proposed GP priors with metric learning can be a future direction. Moreover, in addition to the 3D vision tasks presented in Chapter 3, other computer vision tasks like semantic segmentation may also have the potential to benefit from the correlated latent representations to reduce the flicker among predictions, so the proposed GP priors can be extended to other applications.

Chapter 4 introduces two different methods of learning disentangled image representations when there is prior knowledge in the data domain.

When the factors of variation are well-studied, it is beneficial to also disentangle these factors in the latent space to make the latent image representations more controllable and interpretable. In Publication V, the equivariant representation is used to disentangle the viewpoint changes and the identity for the novel view synthesis task. Though the model combines the benefits of both image-based rendering and the direct pixel regression, the prediction for thin structures (*e.g.*, chair legs and electric poles on the road) can still be inaccurate, since the depth estimation for novel views based on a single source image is challenging. Publication VI leverages semantic attribute labels to divide the image representations into attribute-specific subspaces, so the image representations can be more controllable for interactive fashion retrieval. A block-diagonal memory block module also stores prototype attribute representations to support attribute manipulation. The main limitation of the proposed methods is the requirement of the ground truth labels, which is not always feasible. And the missing attribute labels can also cause problems. In Publication VI, the current method uses the semantic consistency loss to implicitly encourage the missing attribute types to have attribute-specific representations that are close to zero vectors. For 'true' missing attribute (*e.g.*, the sleeve length for paints) it could work, but if the missing is caused by label mistakes, then the zero attribute-specific representations are not expected. Learning disentangled image representations in an unsupervised manner [34, 15, 62] could be a future direction to improve the method.

6. Conclusion

The thesis explores several methods to integrate different types of prior knowledge into deep neural networks. Some of presented methods are tailored for specific computer vision problems (*e.g.*, Publication I for multi-view depth estimation and Publication VI for interactive fashion retrieval), and some methods share the similar paradigm (*e.g.*, Publication II, Publication III and Publication IV learn correlated latent representations via Gaussian process priors).

The presented prior knowledge can be regarded as the domain knowledge for the tasks. More specifically, in this thesis, the prior knowledge can be divided into three categories: (i) the well-studied epipolar geometry, (ii) the correlation introduced by camera poses, and (iii) the known factors of variation. Experimental results show that all prior knowledge improves the performance of deep models for the tasks discussed in the thesis.

For dense 3D reconstruction, Publication I uses the depth distribution to guide the plane sampling for the plane sweep algorithm and utilizes multi-view geometry to extract multiplane representations. The discretized plane volume encodes the scene geometry, and the learned multiplane representation captures the continuous volume opacity.

For applications with given camera motion information (*e.g.*, multi-view stereo and novel view synthesis), the closeness between the camera poses reveals the similarity of the image, which can be exploited to constrain the image representations. To encode prior introduced by camera poses, Publication II, Publication III and Publication IV exploit the latent non-parametric fusion via Gaussian processes. The proposed covariance functions for SE(3) poses enrich the library of GP priors. The use of Gaussian processes in the latent space for prior encoding bridges the gap between computer vision and nonparametric inference, advocating for inter-frame reasoning in more computer vision tasks when the input images are highly correlated.

With known factors of variation, learning disentangled representations makes the deep models more flexible and controllable. For factors like semantic attributes, Publication VI shows that the attribute-specific sub-

spaces can be used to factorize the image representations. For 3D vision tasks like novel view synthesis, it is natural to disentangle the appearance and the viewpoints, and the equivariant representations provide a way of preserving the structure of transformations. Publication V uses the equivariant representation to generate pixels for novel views. To alleviate the limitation of the compact representations and benefit from explicit correspondences, the depth-guided skip connections are proposed to maintain low-level details in synthesis results.

Ultimately, all methods utilize the prior knowledge to regularize the latent space, which makes the latent representations more interpretable and well-structured.

References

[1] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7708–7717, 2018.

[2] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Efficient multi-attribute similarity learning towards attribute-based fashion search. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1671–1679. IEEE, 2018.

[3] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Attribute manipulation generative adversarial networks for fashion images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 10541–10550, 2019.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[5] Neill D. F. Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 766–779, 2008.

[6] Francesco Paolo Casale, Adrian V Dalca, Luca Saglietti, Jennifer Listgarten, and Nicolo Fusi. Gaussian process prior variational autoencoders. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

[7] Angel Xuan Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems (NIPS)*, 29, 2016.

[9] Xu Chen, Jie Song, and Otmar Hilliges. Monocular neural image based rendering with continuous view control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4090–4100, 2019.

[10] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3001–3011, 2020.

[11] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2524–2534, 2020.

[12] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797, 2018.

[13] Taco S. Cohen and Max Welling. Transformation properties of learned visual representations. In *3rd International Conference on Learning Representations (ICLR)*, 2015.

[14] Robert T Collins. A space-sweep approach to true multi-image matching. In *Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 358–363. IEEE, 1996.

[15] Emily L Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017.

[16] James Diebel. Representing attitude: Euler angles, unit quaternions, and rotation vectors. *Matrix*, 58(15-16):1–35, 2006.

[17] Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, and Ping Luo. Camnet: Coarse-to-fine retrieval for camera re-localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2871–2880, 2019.

[18] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15324–15333, June 2021.

[19] Roy Featherstone. *Rigid Body Dynamics Algorithms*. Springer, New York, 2014.

[20] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5515–5524, 2016.

[21] Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski. Towards internet-scale multi-view stereo. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1434–1441, 2010.

[22] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 270–279, 2017.

[23] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3828–3838, 2019.

[24] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

[25] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1735–1742. IEEE, 2006.

[26] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[28] Ari Heljakka, Yuxin Hou, Juho Kannala, and Arno Solin. Deep automodulators. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:13702–13713, 2020.

[29] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[30] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations (ICLR)*, 2017.

[31] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks (ICANN)*, pages 44–51. Springer, 2011.

[32] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[33] Heiko Hirschmuller and Daniel Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1582–1599, 2008.

[34] Qiyang Hu, Attila Szabó, Tiziano Portenier, Paolo Favaro, and Matthias Zwicker. Disentangling factors of variation by mixing them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3399–3407, 2018.

[35] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. DeepMVS: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2821–2830, 2018.

[36] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. DPSnet: End-to-end deep plane sweep stereo. In *International Conference on Learning Representations (ICLR)*, 2019.

[37] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in Neural Information Processing Systems (NIPS)*, 28, 2015.

[38] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[39] Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VSR Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 805–820, 2018.

[40] Dinghuang Ji, Junghyun Kwon, Max McFarland, and Silvio Savarese. Deep view morphing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2155–2163, 2017.

[41] Sing Bing Kang, Richard Szeliski, and Jinxiang Chai. Handling occlusions in dense multi-view stereo. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 103–110. IEEE, 2001.

[42] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[43] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020.

[44] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[45] Marek Kowalski, Stephan J Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. Config: Controllable neural face image generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 299–315. Springer, 2020.

[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 25, 2012.

[47] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. *Advances in Neural Information Processing Systems (NIPS)*, 28, 2015.

[48] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 307–314 vol.1, 1999.

[49] Jeong-gi Kwak, David K Han, and Hanseok Ko. CAFE-GAN: arbitrary face attribute editing with complementary attention feature. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 524–540, 2020.

[50] Karel Lenc and Andrea Vedaldi. Learning covariant feature detectors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 100–117, 2016.

[51] Maxime Lhuillier and Long Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):418–433, 2005.

[52] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. ManiGAN: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7880–7889, 2020.

[53] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016.

[54] William R Mark, Leonard McMillan, and Gary Bishop. Post-rendering 3d warping. In *Proceedings of the 1997 symposium on Interactive 3D graphics*, pages 7–ff, 1997.

[55] Michael F Mathieu, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. *Advances in Neural Information Processing Systems (NIPS)*, 29, 2016.

[56] Claudio Mazzotti, Nicola Sancisi, and Vincenzo Parenti-Castelli. A measure of the distance between two rigid-body poses based on the use of platonic solids. In *Symposium on Robot Design, Dynamics and Control*, pages 81–89. Springer, 2016.

[57] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421, 2020.

[58] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[59] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7588–7597, 2019.

[60] Kyle Olszewski, Sergey Tulyakov, Oliver Woodford, Hao Li, and Linjie Luo. Transformable bottleneck networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7648–7657, 2019.

[61] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3500–3509, 2017.

[62] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.

[63] Vaishakh Patil, Wouter Van Gansbeke, Dengxin Dai, and Luc Van Gool. Don't forget the past: Recurrent depth estimation from monocular video. *IEEE Robotics and Automation Letters*, 5(4):6813–6820, 2020.

[64] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N Metaxas, and Manmohan Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1623–1632, 2017.

[65] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018.

[66] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[67] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. *Advances in Neural Information Processing Systems (NIPS)*, 28, 2015.

[68] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. Disentangling factors of variation for facial expression recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 808–822, 2012.

[69] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015.

[70] Amrita Saha, Megha Nawhal, Mitesh M Khapra, and Vikas C Raykar. Learning disentangled multimodal representations for the fashion domain. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 557–566. IEEE, 2018.

[71] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 501–518, 2016.

[72] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[73] Steven M. Seitz and Charles R. Dyer. Photorealistic scene reconstruction by voxel coloring. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1067–1073, 1997.

[74] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 231–242, 1998.

[75] Arno Solin, Santiago Cortés, Esa Rahtu, and Juho Kannala. Pivo: Probabilistic inertial-visual odometry for occlusion-robust navigation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 616–625. IEEE, 2018.

[76] Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 155–171, 2018.

[77] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 322–337, 2016.

[78] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23(5):903–920, 2012.

[79] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1415–1424, 2017.

[80] Ben Usman, Nick Dufour, Kate Saenko, and Chris Bregler. PuppetGAN: Cross-domain image manipulation by demonstration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9450–9458, 2019.

[81] Kaixuan Wang and Shaojie Shen. MVDepthNet: Real-time multiview depth estimation neural network. In *International Conference on 3D Vision (3DV)*, pages 248–257. IEEE, 2018.

[82] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1164–1174, 2021.

[83] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. NerfingMVS: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5610–5619, 2021.

[84] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Interpretable transformations with encoder-decoder networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5726–5735, 2017.

[85] Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. *Advances in Neural Information Processing Systems (NIPS)*, 28, 2015.

[86] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[87] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5534, 2019.

[88] Jianfu Zhang, Yuanyuan Huang, Yaoyi Li, Weijie Zhao, and Liqing Zhang. Multi-attribute transfer via disentangled representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9195–9202, 2019.

[89] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1520–1528, 2017.

[90] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian-sheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4913–4922, 2019.

[91] Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. DeepTAM: Deep tracking and mapping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[92] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858, 2017.

[93] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018.

[94] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301. Springer, 2016.

[95] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Joshua B. Tenenbaum, and William T. Freeman. Visual object networks: Image generation with disentangled 3D representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Aalto-DT 194/2022

ISBN 978-952-64-1072-2 (printed)
ISBN 978-952-64-1073-9 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL THESES