



Universiteit  
Leiden  
The Netherlands

## Exploring deep learning for intelligent image retrieval

Chen, W.

### Citation

Chen, W. (2021, October 13). *Exploring deep learning for intelligent image retrieval*. Retrieved from <https://hdl.handle.net/1887/3217054>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3217054>

**Note:** To cite this publication please use the final published version (if applicable).

# Exploring Deep Learning for Intelligent Image Retrieval

Wei Chen

Copyright © 2021 Wei Chen, All Rights Reserved

ISBN 978-94-6419-293-3

Printed by Gildeprint, The Netherlands

Cover design: Zhihan Zhao, Wei Chen

# Exploring Deep Learning for Intelligent Image Retrieval

**Proefschrift**

ter verkrijging van  
de graad van doctor aan de Universiteit Leiden,  
op gezag van rector magnificus prof.dr.ir. H. Bijl,  
volgens besluit van het college voor promoties  
te verdedigen op woensdag 13 oktober 2021  
klokke 15.00 uur

door

**Wei Chen**

geboren te Guizhou, China  
in 1991

## Promotiecommissie

Promotors: Prof. dr. M.S. Lew  
Prof. dr. A. Plaat

Overige leden: Prof. dr. T.S. Chua (National University of Singapore)  
Prof. dr. B.P.F. Lelieveldt  
Prof. dr. T.H.W. Bäck  
Dr. E.M. Bakker  
Dr. K.J. Wolstencroft



Wei Chen was financially supported through the China Scholarship Council (CSC) to participate in the PhD programme of Leiden University. Grant number 201703170183.

The research in this thesis was performed at the LIACS MediaLab, Leiden University, The Netherlands, and we would like to thank the NVIDIA Corporation for the donation of GPU cards.

*To my family*



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background . . . . .	2
1.2	Research Questions and Contributions . . . . .	3
1.3	Thesis Overview . . . . .	5
<b>2</b>	<b>A Comprehensive Review of Deep Image Retrieval</b>	<b>9</b>
2.1	Introduction . . . . .	10
2.1.1	Summary of progress since 2012 . . . . .	12
2.1.2	Key challenges . . . . .	14
2.2	Deep Convolutional Neural Networks . . . . .	15
2.2.1	A brief introduction to deep learning . . . . .	15
2.2.2	Popular backbone DCNN architectures . . . . .	16
2.3	Retrieval with Off-the-Shelf DCNN Models . . . . .	17
2.3.1	Deep feature extraction . . . . .	18
2.3.1.1	Network feedforward scheme . . . . .	18
2.3.1.2	Deep feature selection . . . . .	19
2.3.1.3	Feature fusion strategy . . . . .	21
2.3.2	Deep feature enhancement . . . . .	23
2.3.2.1	Feature aggregation . . . . .	23
2.3.2.2	Feature embedding . . . . .	23
2.3.2.3	Attention mechanisms . . . . .	26
2.3.2.4	Deep hash embedding . . . . .	27
2.4	Retrieval via Learning DCNN Representations . . . . .	28
2.4.1	Supervised fine-tuning . . . . .	29
2.4.1.1	Classification-based fine-tuning . . . . .	29
2.4.1.2	Verification-based fine-tuning . . . . .	30
2.4.2	Unsupervised fine-tuning . . . . .	34
2.4.2.1	Mining samples with manifold learning . . . . .	35
2.4.2.2	AutoEncoder-based frameworks . . . . .	36
2.5	State of the Art Performance . . . . .	38
2.5.1	Datasets . . . . .	38
2.5.2	Evaluation metrics . . . . .	38
2.5.3	Performance comparison and analysis . . . . .	39



2.6	Chapter Conclusions . . . . .	43
<b>3</b>	<b>Domain Uncertainty based on Information Theory for Cross-modal Hash Retrieval</b>	<b>47</b>
3.1	Introduction . . . . .	48
3.2	Cross-modal Hash Learning . . . . .	49
3.3	Domain Uncertainty Measurement via Information Theory . . . . .	50
3.3.1	Information theory and domain uncertainty . . . . .	50
3.3.2	Multi-level feature preserving . . . . .	51
3.3.3	Classification-based objective function . . . . .	53
3.4	Experiments and Evaluations . . . . .	54
3.4.1	Implementation details . . . . .	54
3.4.2	Datasets . . . . .	54
3.4.3	Performance and evaluation . . . . .	54
3.5	Chapter Conclusions . . . . .	57
<b>4</b>	<b>Integrating Information Theory and Adversarial Learning for Cross-modal Retrieval</b>	<b>59</b>
4.1	Introduction . . . . .	60
4.2	Related Work . . . . .	61
4.2.1	Cross-modal representation learning and matching . . . . .	61
4.2.2	Adversarial learning for cross-modal retrieval . . . . .	62
4.2.3	Information-theoretical feature learning . . . . .	63
4.3	Proposed Approach . . . . .	64
4.3.1	Problem formulation . . . . .	64
4.3.2	Integrating information theory & adversarial learning . . . . .	64
4.3.2.1	Information entropy and modality uncertainty . . . . .	64
4.3.2.2	Adversarial learning and information entropy . . . . .	65
4.3.3	KL-divergence for cross-modal feature projection . . . . .	66
4.4	Implementation and optimization . . . . .	67
4.4.1	Combining information theory & adversarial learning . . . . .	67
4.4.2	KL-divergence for similarity preserving . . . . .	69
4.4.3	Instance label classification . . . . .	70
4.4.3.1	Categorical cross-entropy loss . . . . .	70
4.4.3.2	KL-divergence for data imbalance . . . . .	70
4.4.4	Bi-directional triplet constraint . . . . .	71
4.5	Experiments . . . . .	73
4.5.1	Datasets and settings . . . . .	73
4.5.2	Performance evaluation . . . . .	74
4.5.2.1	Results on the Flickr30K and MS-COCO datasets . . . . .	74
4.5.2.2	Results on CUHK-PEDES dataset . . . . .	75
4.5.2.3	Results on Flickr8K dataset . . . . .	76
4.5.3	Ablation study . . . . .	76

4.5.3.1	Analysis of KL-divergence for data imbalance . . . . .	77
4.5.3.2	Analysis of KL-divergence for cross-modal feature projection . . . . .	77
4.5.3.3	Analysis of adversary combining . . . . .	77
4.5.3.4	Analysis of temperature $\tau$ . . . . .	78
4.5.3.5	Distribution visualization . . . . .	78
4.5.3.6	Analysis of complexity and stability . . . . .	80
4.5.4	Further exploring . . . . .	81
4.6	Chapter Conclusions . . . . .	83
<b>5</b>	<b>On the Exploration of Incremental Learning for Fine-grained Image Retrieval</b>	<b>85</b>
5.1	Introduction . . . . .	86
5.2	Related Work . . . . .	87
5.3	Proposed Approach . . . . .	87
5.3.1	Semantic preserving loss . . . . .	88
5.3.2	Knowledge distillation loss . . . . .	89
5.3.3	Maximum mean discrepancy loss . . . . .	89
5.4	Experiments . . . . .	91
5.4.1	Datasets and experimental settings . . . . .	91
5.4.2	One-step incremental learning for FGIR . . . . .	91
5.4.3	Multi-step incremental learning for FGIR . . . . .	94
5.4.4	Validation with image classification . . . . .	97
5.4.5	Training time comparison . . . . .	98
5.4.6	Components analysis . . . . .	98
5.5	Chapter Conclusions . . . . .	98
<b>6</b>	<b>Feature Estimations based Correlation Distillation for Incremental Image Retrieval</b>	<b>101</b>
6.1	Introduction . . . . .	102
6.2	Related Work . . . . .	103
6.3	Correlations Distillation for Incremental Image Retrieval . . . . .	104
6.3.1	Problem formulation . . . . .	104
6.3.2	Correlations distillation for one-task scenario . . . . .	106
6.3.3	Feature estimation for multi-task scenario . . . . .	106
6.4	Experiments . . . . .	110
6.4.1	Datasets and experimental setup . . . . .	110
6.4.2	One-task scenario evaluation . . . . .	111
6.4.3	Multi-task scenario evaluation . . . . .	114
6.4.4	Ablation study . . . . .	116
6.4.5	Retrieval visualization . . . . .	120
6.5	Chapter Conclusions . . . . .	121

<b>7</b>	<b>Lifelong Image Retrieval via Dual Knowledge Distillation</b>	<b>125</b>
7.1	Introduction . . . . .	126
7.2	Related Work . . . . .	127
7.3	The Lifelong Image Retrieval Problem . . . . .	128
7.4	Dual Knowledge Distillation . . . . .	129
7.4.1	Knowledge distillation by frozen teacher . . . . .	129
7.4.2	Representative data generation . . . . .	130
7.4.3	Self-motivated learning on the mixed data . . . . .	131
7.4.4	Auxiliary distillation by on-the-fly teacher . . . . .	131
7.5	Experiments . . . . .	132
7.5.1	Dataset splits . . . . .	132
7.5.2	Training details . . . . .	133
7.5.3	Performance evaluation . . . . .	133
7.5.4	Further explorations . . . . .	141
7.6	Chapter Conclusions . . . . .	142
<b>8</b>	<b>New Ideas and Trends in Deep Multimodal Content Understanding</b>	<b>143</b>
8.1	Introduction . . . . .	144
8.2	Multimodal Applications . . . . .	145
8.2.1	Uni-directional applications . . . . .	145
8.2.2	Bi-directional applications . . . . .	146
8.3	Recent Advances in Content Understanding . . . . .	148
8.3.1	Deep multimodal structures . . . . .	148
8.3.2	Multimodal feature extraction . . . . .	151
8.3.3	Common latent space learning . . . . .	156
8.4	Results and Discussions . . . . .	162
8.5	Chapter Conclusions . . . . .	168
<b>9</b>	<b>Conclusions</b>	<b>169</b>
9.1	Limitations and Possible Solutions . . . . .	172
9.2	Future Research Directions . . . . .	172
	<b>Bibliography</b>	<b>175</b>
	<b>List of Abbreviations</b>	<b>195</b>
	<b>English Summary</b>	<b>197</b>
	<b>Nederlandse Samenvatting</b>	<b>199</b>
	<b>Curriculum Vitae</b>	<b>205</b>