Privileged Machine Learning for Prediction

Yangyang Shu

Supervisor: Dr. Guandong Xu

Advanced Analytics Institute
Faculty of Engineering and Information Technology
University of Technology Sydney

A thesis submitted for the degree of PhD
October 2021
© 2021 by Yangyang Shu
All Rights Reserved

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Yangyang Shu declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. In addition, this document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed
prior to publication.

11 October 2021

Acknowledgments

I would like to express my sincere gratitude towards everyone who made this thesis possible.

First and foremost, I would like to show my deepest gratitude to my supervisor, Prof. Guandong Xu, a respectable, responsible and resourceful scholar, who has provided me with valuable guidance in every stage of the writing of this thesis. I have learned a lot from him about what research is, and the patience he has shown towards me is genuinely exceptional. Without his enlightening instruction, impressive kindness and patience, I could not have completed my thesis. His keen and vigorous academic observation enlightens me not only in this thesis but also in my future study. I could not have imagined a better supervisor.

I would especially like to thank my Cooperating teachers Dr. Yulei Sui and Dr. Hongyu Zhang for helping me to grow as a researcher. I truly admire their passion, drive, and ability to express complicated things in the most straightforward way possible. They have continuously challenged me and has been a great source of motivation and advice. I sincerely appreciate all the efforts they have taken to shape me as a person.

I am grateful to my co-supervisor Dr. Qian Li and Dr. Shaowu Liu for many insightful ideas they provided when I was stuck. I am also grateful to all the members of the UTS Data Science and Machine Intelligence Lab for all the good times we had both during and outside research. I express my gratitude towards friends I made.

I am thankful to UTS for providing financial and technical support for my research.

I would also like to thank Dr. Baosheng Yu who has helped me to develop the fundamental and essential academic competence. My sincere appreciation also goes to Dr. Chang Xu, the University of Sydney, who helps my research with great cooperation.

I can not thank my girlfriend Yuying, enough for her sacrifice, support, advice, and live throughout this journey. She is the reason I smile despite when chips are down.

Finally, and most importantly, I would like to thank my parents. I could not have done this without their guidance, encouragement, selfless love and support.

Abstract

Machine learning for prediction suffers from asymmetric distribution, such as posterior information, future information and hidden information. With some additional information only available in training, how to learn a machine learning model with them remains a key challenge. Despite recent advances in important domains such as vision and medicine, the standard learning under privileged information paradigm does not offer a satisfactory solution for learning variational privileged information. In this thesis, I will introduce how to learn under variational privileged information by leveraging asymmetric distribution and machine learning algorithms, specifically via 1) semi-supervised learning, 2) adversarial learning, 3) multi-task learning, 4) slack variable learning in support vector regression. We evaluate the proposed method in three applications: photo aesthetic assessment enhanced by high-level aesthetic attributes hidden in photos; music emotion recognition from songs with the help of implicit information about music elements and musical styles judged by composers; and multiple object recognition from images with the help of implicit information about the object's importance conveyed by the list of manually annotated image tags. Experiment results demonstrate that the proposed methods are superior to the classic learning paradigm when solving practical problems. In summary, in this thesis, we propose privileged machine learning for prediction. The detailed contents are follow:

1. propose a unified framework to systematically address the aforementioned three forms of privileged information. The proposed V-SVR+ method integrates continuous, ordinal, and binary PI into the learning process of support vector regression (SVR) via three losses. For continuous privileged information, we define a linear correcting (slack) function in the privileged information space to estimate slack variables in the standard SVR method using privileged information. For the ordinal relations of privileged information, we first rank the privileged information and then, regard this ordinal privileged information as auxiliary information used in the learning process of the SVR model. For the binary or Boolean privileged information, we infer a probabilistic dependency between the privileged information and labels from the summarized privileged information knowledge. Then, we transfer the privileged information knowledge to constraints and form a constrained optimization problem.

2. propose a novel approach of photo aesthetic assessment under the help of

aesthetic attributes. The aesthetic attributes are used as privileged information (PI), which is often available during training phase but unavailable in prediction phase due to the high collection expense. The proposed framework consists of a deep multi-task network as generator and a fully connected network as discriminator. Deep multi-task network learns the aesthetic attributes and score simultaneously to capture their dependencies and extract better feature representations. Specifically, we use ranking constraint in the label space, similarity constraint and prior probabilities loss in the privileged information space to make the output of multi-task network converge to that of ground truth. Adversarial loss is used to identify and distinguish the predicted privileged information of deep multi-task network from the ground truth PI distribution. Experimental results on two benchmark databases demonstrate the superiority of the proposed method to state-of-the-art.

3. propose a novel privileged learning based framework that fully explores the musical domain knowledge thus to enhance the emotion recognition. Particularly, to best of our knowledge, we are the first to construct a systematic taxonomy of the dimension and style-related music information by applying a domain knowledge of musicology and psychology. Then, we customize Restricted Boltzmann Machine (RBM) to generate the informative feature representation that captures the intrinsic dependency between musical elements and music styles. Finally, we formulate the generated feature representation as privileged information (PI), and develop a PI-based support vector regression (i.e., SVR+) for music emotion recognition task. Extensive experimental results on two benchmark databases demonstrate the superior performance on emotion recognition compared against the state-of-the-art baselines.

4. propose a novel regression algorithm via ordinal Privileged Information, which takes into consideration of ordinal form of privileged information. By integrating ordinal constraint into the learning process, the privileged information and the dependencies in images features can enhance the object recognition. In optimization, we adopt maximum margin regression model. Alternating Direction Method of Multipliers (ADMM) is developed to optimize this proposed model. We evaluate the proposed method on multiple object recognition from images with the help of implicit information about the object's importance conveyed by the list of manually annotated image tags. Experimental results demonstrate that the proposed method can effectively take the advantage of ordinal privileged information.

Publications

The majority of the thesis has been published in peer-reviewed conference and journal proceedings.

Journal Publications

- **Yangyang Shu**, Qian Li and Guandong Xu. "Privileged Multi-Task Learning for Attribute-Aware Aesthetic Assessment." *Pattern Recognition*, 2021. *Under review*.
- **Yangyang Shu**, Qian Li and Guandong Xu. "Semi-Supervised Adversarial Learning for Attribute-Aware Photo Aesthetic Assessment." *IEEE Transactions on Multimedia (TMM)*, 2021. In press, 10.1109/TMM.2021.3117709
- **Yangyang Shu**, Qian Li, Chang Xu, Shaowu Liu and Guandong Xu. "V-SVR+: Support Vector Regression with Variational Privileged Information." *IEEE Transactions on Multimedia (TMM)*, 2021. In press, 10.1109/TMM.2021.3060955
- **Yangyang Shu** and Guandong Xu. "Learning with Privileged Information for Music Emotion Recognition." *IEEE Transactions on Affective Computing (TAC)*, 2021. *Under review*.
- **Yangyang Shu**, Qian Li, Shaowu Liu and Guandong Xu. "Learning with privileged information for photo aesthetic assessment." *Neurocomputing*, 2020, 404, pp.304-316.
- Shangfei Wang, Can Wang, Tanfang Chen, Yaxin Wang, **Yangyang Shu** and Qiang Ji. "Video affective content analysis by exploring domain knowledge." *IEEE Transactions on Affective Computing (TAC)*, 2019. 10.1109/TAFFC.2019.2912377

Conference Publications

- **Yangyang Shu**, Yulei Sui, Hongyu Zhang and Guandong Xu. "Perf-AL: Performance Prediction for Configurable Software through Adversarial Learning" *Empirical Software Engineering and Measurement (ESEM)*, Bari, Italy, Ictorber 5-7, 2020, pp. 1-11.

- **Yangyang Shu** and Guandong Xu. "Emotion Recognition from Music Enhanced by Domain Knowledge." *Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, Yanuca Island, Fiji, AUgust 26-30, 2019, pp. 121-134.
- **Yangyang Shu** and Shangfei Wang. "Emotion recognition through integrating EEG and peripheral signals." *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, March 5-9, 2017, pp. 2871-2875.

Contents

Acknowledgments **2**

Abstract **3**

Publications **5**

1 Introduction **1**
1.1 Background . . . 1
1.2 Existing Efforts and Limitations . . . 1
1.3 Insights and Our Solution . . . 3
1.4 Thesis Overview . . . 3
1.5 Key Contributions and Impact . . . 5

2 Related Work **7**
2.1 Learning Under Privileged Information . . . 7
2.2 Attribute-Aware Photo Aesthetic Assessment . . . 8
2.2.1 Low-Level Attributes . . . 8
2.2.2 Deep Features . . . 9
2.2.3 High-Level Attributes . . . 9
2.2.3.1 Fully attribute-annotated photos . . . 9
2.2.3.2 Partially attribute-annotated photos . . . 10
2.3 Music Emotion Recognition . . . 10
2.3.1 Low-level features . . . 10
2.3.2 Deep features . . . 11
2.3.3 High-level features . . . 11
2.4 Multiple Object Recognition . . . 12
2.4.1 Object Recognition with Traditional Machine Learning Method . . . 12
2.4.2 Object Recognition with Deep Learning Method . . . 12
2.5 Discussion of Privileged Machine Learning . . . 13

3 Preliminaries **14**
3.1 Support Vector Regression . . . 14
3.2 Generative Adversarial Networks . . . 14

3.3 Semi-supervised Adversarial Learning . . . 15
3.4 Multi-task Learning . . . 16

4 V-SVR+: Support Vector Regression with Variational Privileged Information 17
4.1 Introduction . . . 17
4.2 Methodology . . . 18
4.2.1 SVR with Continuous Privileged Information . . . 19
4.2.2 SVR with Ordinal Privileged Information . . . 21
4.2.3 SVR with Binary Privileged Information . . . 24
4.3 Conclusion . . . 25

5 Learning with Privileged Information for Photo Aesthetic Assessment 27
5.1 Introduction . . . 27
5.2 Preliminaries . . . 29
5.3 Methodology . . . 29
5.3.1 Full Attribute-annotated Photos: Privileged Multi-task Learning for Attribute-aware Aesthetic Assessment . . . 29
5.3.1.1 Basic Network . . . 30
5.3.1.2 Guidance in Label Space . . . 31
5.3.1.3 Guidance in Privileged Information Space . . . 31
5.3.2 Partially Attribute-annotated Photos: Semi-supervised Adversarial Learning for Attribute-aware Photo Aesthetic Assessment 34
5.3.2.1 Adversarial Loss . . . 36
5.3.2.2 Supervised Loss . . . 36
5.3.2.3 Reconstruction Loss . . . 37
5.3.2.4 The Learning Objectives of Each Network . . . 37
5.4 Experiments . . . 39
5.4.1 Experimental Conditions . . . 39
5.4.1.1 Datasets . . . 39
5.4.1.2 Implementation Details . . . 40
5.4.1.3 Evaluation Metric . . . 41
5.4.1.4 Baselines . . . 41
5.4.1.5 Experiment Design on Full Attribute-annotated Photo Aesthetic Assessment . . . 43
5.4.1.6 Experiment Design on Partially Attribute-annotated Photo Aesthetic Assessment . . . 43
5.4.2 Results of Within-Database Experiments . . . 43

5.4.2.1 Fully Attribute-annotated Results . . . 43
5.4.2.2 Partially Attribute-annotated Results . . . 44
5.4.3 Results of Cross-Database Experiments . . . 47
5.4.3.1 Fully Attribute-annotated Results . . . 47
5.4.3.2 Partially Attribute-annotated Results . . . 47
5.5 Discussion and Conclusion . . . 49

6 Learning with Privileged Information for Music Emotion Recognition 50
6.1 Introduction . . . 50
6.2 Summary of Domain Knowledge . . . 53
6.2.1 Dimension-related Domain Knowledge . . . 53
6.2.1.1 Rhythm . . . 53
6.2.1.2 Tonality . . . 54
6.2.1.3 Timbre . . . 54
6.2.1.4 Dynamics . . . 55
6.2.1.5 Pitch . . . 55
6.2.2 Style-related Domain Knowledge . . . 55
6.3 Methodology . . . 57
6.3.1 Problem Statement . . . 58
6.3.2 Modeling Relations between Musical Elements and Music Styles by RBM . . . 58
6.3.3 Proposed LUPI for Music Emotion Recognition . . . 60
6.3.4 Optimization . . . 65
6.4 Experiments . . . 65
6.4.1 Experimental Conditions . . . 65
6.4.1.1 Datasets . . . 65
6.4.1.2 Privileged Information . . . 66
6.4.1.3 Evaluation Details . . . 67
6.4.1.4 Baselines . . . 67
6.4.2 Comparison Results . . . 69
6.4.3 Analysis of Fusion Methods . . . 70
6.4.4 Comparison with Different Sorts of Privileged Information Features . . . 71
6.4.5 Analysis the Effectiveness of Privileged Information . . . 73
6.5 Conclusion . . . 73
6.5.1 Summary . . . 73
6.5.2 Future Work . . . 74

7 Learning Ordinal Privileged Information for Multiple Object Recognition 75
7.1 Introduction . . . 75
7.2 Methodology . . . 77
7.2.1 Learning with Ordinal Privileged Information . . . 77
7.2.2 Optimization of the Proposed Method . . . 79
7.3 Experiments . . . 81
7.3.1 Experimental Conditions . . . 81
7.3.1.1 Dataset . . . 81
7.3.1.2 Experimental Design . . . 82
7.3.1.3 Performance Metrics . . . 82
7.3.2 Experimental Results and Analysis . . . 82
7.3.3 Comparison to Related Works . . . 83
7.4 Conclusion . . . 83

8 Conclusion 84
8.1 Summary . . . 84
8.2 Future work . . . 85

List of Figures

1.1 Examples of continuous, ordinal and binary form of privileged information. 2

4.1 The framework of the proposed V-SVR+ model, consisting of three modules to manage three forms of PI respectively. The input of $\{C, O, B\}$ represents continuous, ordinal and binary forms of privileged information. w_1, b_1, w_2, b_2, w_3, b_3 are their parameters. 18

5.1 Different aesthetic qualities w.r.t. different photo attributes as privileged information. The first and third rows include photos with high aesthetic qualities. The second and fourth rows include photos with low aesthetic qualities. 28

5.2 The framework of the proposed method. In training phase, two photos are first trained respectively. Utilizing pairwise photos aims to learn ranking loss functions. In our generator networks, $d^* + 1$ tasks correspond to d^* learning tasks for aesthetic attributes and one task for predicting the overall score. These aesthetic attributes enable to provide more detailed feedback on automatic aesthetic assessment. "positive" and "negative" refer to coexistent and exclusive relations among multiple tasks. 30

5.3 The framework of the proposed semi-supervised adversarial learning. . 36

5.4 Examples of predicted aesthetic scores on the AADB database: RA-DCNN/ATTGAN/Ours/Ground truth. 45

5.5 Examples of predicted aesthetic scores on the AVA database: RA-DCNN/ATTGAN/Ours/Ground truth. 45

6.1 Example of music piece with prior information i.e. musical elements and music styles. This music piece is summarized from "The Butterfly Lovers", which is a classical music. From this music score, we can learn main music information, e.g. "Andante" represents musical tempo; "mp" represents musical loudness; Treble clef and bass clef indicate musical pitch; Sharp and flat refer to musical mode (See Figure 6.2); Different instruments result in different musical brightness [Saitis and Siedenburg, 2020]. These musical information collected in a laboratory environment are beneficial for MER, which can be used in training phase to improve the performance of model. 51
6.2 The major mode and minor mode of music corresponding to stave and musical mark. 54
6.3 The framework of the proposed method. 57
6.4 The distributions of all music styles on both MediaEval and AMG 1608 databases. The X-axis is the different music styles and the Y-axis is their percentage in the corresponding databases. 67

7.1 Examples of different privileged information forms 76
7.2 Example PASCAL images with relative importance of the objects collected on Mechanical Turk. 76

List of Tables

5.1 Attribute annotations on the AADB and AVA database. 39
5.2 Experimental results on the AADB and AVA databases. 43
5.3 Within database experimental results of semi-supervised adversarial learning for aesthetic assessment with four missing rates of attribute-annotated photos on the AADB and AVA databases. 44
5.4 Comparison to state-of-the-art attribute-aware methods on the AADB and AVA database. 46
5.5 Comparison to state-of-the-art attribute-unaware methods on the AADB and AVA databases. 46
5.6 SRCC (Spearman Rank-order Correlation Coefficient) for training and testing on Cross-Database. 47
5.7 Cross-database evaluation of semi-supervised adversarial learning for aesthetic assessment with four missing rates of attribute-annotated photos. 48
5.8 Cross-database evaluation of SRCC reported on current studies. 48

6.1 Analysis for emotional effects on musical elements and music styles. Note that High-A, Low-A, Positive-V and Negative-V represent high arousal, low arousal, positive valence and negative valence, respectively. $\surd$ denotes the existence of high dependency. $\uparrow$ denotes the associated emotions under improving the row status, $\downarrow$ denotes the associated emotions under descending the row status. 53
6.2 Scale of dynamic markings. 55
6.3 Comparison with the related works on the MediaEval 2015 database. . . 69
6.4 Comparison with the related works on the AMG 1608 database. 69
6.5 Emotion recognition on MediaEval 2015 database. 70
6.6 Emotion recognition on AMG 1608 database. 70
6.7 Exploring musical element-related privileged information for emotion recognition from songs on both the MediaEval 2015 and AMG 1608 databases. "-/-" denotes RMSE and R respectively, "total" represents the all musical elements. 71

6.8 Exploring music style-related privileged information for emotion recognition from songs on both the MediaEval and AMG 1608 databases. "-/-" denotes RMSE and R respectively, "total" represents all the music styles. 72
6.9 The performance of the P-values of the proposed SVR+ with musical elements, music styles and privileged information of RBM compared with the method which ignores privileged information. "-/-" denotes P-values of RMSE and R respectively. 72

7.1 Average precision (AP) scores per-class results on the Pascal VOC 2007 database. The results of the proposed method rank first in 11 out of 20 classes. 82

Chapter 1

Introduction

1.1 Background

In most machine learning for prediction, symmetric information distribution is used to both train and test the model. However, one can access not only the input/output training pairs of the task we want to learn but also additional information on the training examples (the so-called privileged information (PI)), which can include posterior information (e.g. the post-classification diagnosis of a physician, in a medical setting), future information (e.g. future information about the price of a stock, in a stock-price prediction setting) or hidden information (e.g. hidden information about the importance of objects in multiple object recognition). Typically, this additional information is more informative to enhance learning than using training data alone; thus, how to learn a machine learning model with privileged information becomes more and more important. Although current studies propose some advanced techniques in important domains such as vision and medicine, the standard learning under privileged information paradigm does not offer a satisfactory solution to learning variational privileged information and the obtaining of PI is typically not easily because of higher data collection costs and/or a lack of access to domain expertise.

1.2 Existing Efforts and Limitations

Over years, many machine learning or deep learning methods with privileged information have been proposed. For example, Vapnik and Vashist et al. [Vapnik and Vashist, 2009] is the first to propose the concept of privileged information. Then, they give the *learning using privileged information (LUPI)* paradigm where privileged information is used to predict the slack variables in the support vector machine. Because SVM+ is computationally expensive, various works of the optimization algorithm have been proposed to solve SVM+, such as extended L1-norm SVM [Niu and Wu, 2012], L2-norm SVM+[Li et al., 2016a] and W-SVM [Lapin et al., 2014]. Thus, priv-

ileged information has been widely applied in various machine learning tasks, e.g. classification and regression [Lambert et al., 2018], computer vision [Xu et al., 2015] [Sharmanska et al., 2013] and multi-label learning [You et al., 2017] in recent years. However, the above-mentioned approaches only consider the single form of privileged information, and machine learning for prediction often encounters variational privileged information, which is always limited and difficult to solve and exploit.

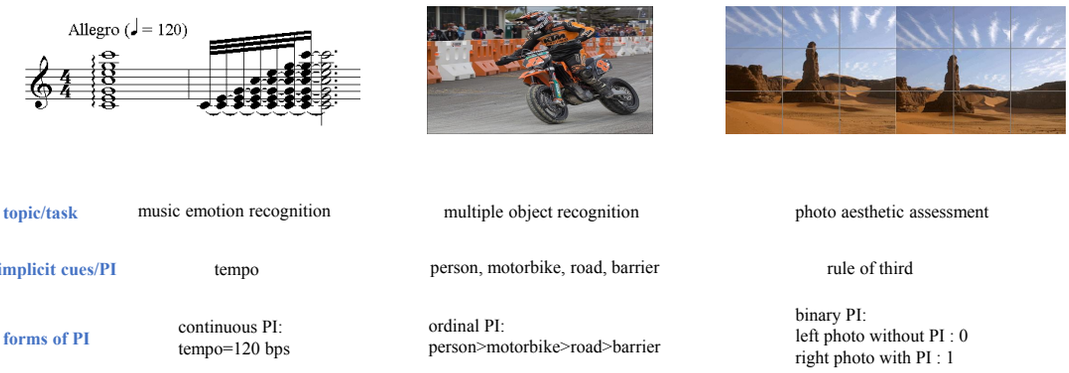


Figure 1.1: Examples of continuous, ordinal and binary form of privileged information.

The key challenge for privileged information for prediction is how to learn a machine learning model with variational additional information only available in training. Most existing algorithms assume that the privileged information is provided in the same form, which is often not true in real-world applications. For example, as shown in Figure 1.1, privileged information exists in different forms. Taking an example of continuous privileged information, tempo, in music pieces are continuous values. It is one of the main musical dimensions hidden music pieces and plays an important role in music emotion recognition. For the example of ordinal privileged information shown in the middle of Figure 1.1, the object importance implicit in the ordinal information can be regarded as ordinal PI to enhance the learning task of object recognition. For the example of binary privileged information, as we can see from the right of Figure 1.1, high-level descriptive such as rule of third explicitly predict possible image cues, which can be used as binary privileged information for photo aesthetic assessment. We denote these attributes as either 0 or 1 to represent the existence or not.

Therefore, considering different forms of PI in training will be of more benefit to learning than using unified privileged information.

1.3 Insights and Our Solution

This thesis aims to learn under variational privileged information by leveraging asymmetric distribution and machine learning algorithms. The proposed methods enhance machine learning for prediction with different forms of privileged information. Particularly, first, we focus on variational privileged information in support vector regression. Specifically, the proposed approach we called v-SVR+ is based on the maximum margin regression model. The proposed v-SVR+ method consists of continuous PI-based SVR, ordinal PI-based SVR and binary PI-based SVR. For continuous PI-based SVR, we apply the Lagrangian multiplier method [Everett III, 1963] to solve this problem. For ordinal PI-based SVR, we solve this problem using the alternating direction method of multipliers (ADMM) [Boyd et al., 2011]. For binary PI-based SVR, we solve this problem using stochastic gradient descent (SGD) due to the easy calculation of the partial derivative of the required parameter. Second, we focus on variational privileged information in the deep learning method where we integrate vairiational privileged information into the learning process of network only in the training phase. We also explore privileged machine learning methods on three different applications. The proposed privileged multi-task framework and semi-supervised adversarial learning framework are used in photo aesthetic assessment with fully attribute-annotated and partially attribute-annotated photos, respectively. For music emotion recognition, we use the implicit information about music elements and musical styles judged by composers as continuous privileged information to improve the performance of machine prediction for prediction. For multiple object recognition, we use the implicit information about object importance as ordinal privileged information to enhance the recognition performance.

1.4 Thesis Overview

This thesis explores the research questions in detail with the following structure.

In Chapter 2, we first provide a background of the learning under privileged information paradigm, with some extending method and optimization such as the l_1 SVM+, l_2 SVM+ and R-SVM+. Then we discuss some example that privileged information can be used to improve the performance of their proposed models. We end the chapter with three applications: photo aesthetic assessment, music emotion recognition and multiple object recognition.

In Chapter 3, we briefly review four popular methods because our proposed methods are used and improved with these machine learning algorithms. They are support vector regression [Drucker et al., 1997], generative adversarial network [Good-

fellow et al., 2014], semi-supervised adversarial learning [Lai et al., 2017] and multi-task learning [Caruana, 1997].

In Chapter 4, we learn the first research question by exploiting variational privileged information in support vector regression. We first start by observing that many regression tasks encounter an asymmetric distribution of information between training and testing phases. Next, our proposed V-SVR+ model, integrates continuous, ordinal, and binary PI into the learning process of support vector regression (SVR) via three losses. In this chapter, we introduce three kinds of privileged information based on the maximum margin regression model and give their optimizations. For continuous privileged information, we define a linear correcting (slack) function in the privileged information space to estimate slack variables in the standard SVR method using privileged information. For the ordinal relations of privileged information, we first rank the privileged information and then, regard this ordinal privileged information as auxiliary information used in the learning process of the SVR model. For the binary or Boolean privileged information, we infer a probabilistic dependency between the privileged information and labels from the summarized privileged information knowledge. Then, we transfer the privileged information knowledge to constraints and form a constrained optimization problem.

In Chapter 5, we tackle the second research question by developing a novel framework that contains two components, a deep multi-task network as generator and a fully connected network as discriminator. High-level attributes as privileged information used in the proposed network where similarity constraint represents continuous PI, ranking constraint represents ordinal PI and prior probabilities loss represents binary PI used as loss function to benefit the aesthetic assessment. Furthermore, we propose novel ideal strategies for photo aesthetic assessment from partially attribute-annotated photos, which can greatly reduce the reliance on manual attribute annotation. For attribute-annotated training data, we reconstruct the photo with aesthetic attributes, score from ground truth, and informative feature representation from the penultimate layer features of score-attributes generator, then generate fake aesthetic attributes and score using the output of the photo generator as the input of the score-attributes generator. For no attribute-annotated training data, we reconstruct the photo with the output of the score-attributes generator and informative feature representation as the input of the photo generator.

In Chapter 6, we address the third research question by presenting a novel method that captures the dependencies between music dimensions and music styles to become new privileged information representation, and further propose the SVR+ model, i.e. learning under new privileged information representation in the standard SVR method of emotion recognition and regression. We first summarize dimension-

related domain knowledge and style-related domain knowledge from musicology and psychology. Then, we customize Restricted Boltzmann Machine (RBM) to generate the informative feature representation that captures the intrinsic dependency between musical elements and music styles. Finally, we formulate the generated feature representation as privileged information (PI), and develop a PI-based support vector regression (i.e., SVR+) for music emotion recognition task.

In Chapter 7, we exploit multiple object recognition from images with the help of implicit information about the object's importance conveyed by the list of manually annotated image tags. In this chapter, we propose to use the object's importance in images as ordinal privileged information to improve the performance of multiple object recognition. We resort to alternating direction method of multipliers [Boyd et al., 2011] to optimize our method.

Finally, we summarize this thesis along with several future directions in Chapter 8.

1.5 Key Contributions and Impact

The main contributions of this thesis include advancements of different forms of privileged information in SVR and deep learning method, applied on photo aesthetic assessment, music emotion recognition, and multiple object recognition.

- **Novel prediction models.** We propose a unified framework that integrates continuous, ordinal, and binary privileged information into the learning process of SVR. To our knowledge, the proposed V-SVR+ algorithm with the corresponding optimization strategies, is the first to learn from variational privileged information.

- **A multi-task network aesthetics rating.** We propose a novel deep multi-task method that incorporates high-level aesthetic attributes as privileged information (PI), from which better feature representations are constructed via multi-task learning.

- **Variational privileged information for aesthetics.** The proposed approach uses similarity constraint as continuous privileged information upon the output layer closing to ground truth , uses ranking constraint to capture the ordinal relation between attributes and aesthetic score, uses prior probabilities loss as binary privileged information to capture the dependencies among multiple attributes, and uses adversarial loss to force the joint distributions of the score

and aesthetic attributes from generative network to converge to the ground-truth. By integrating variational privileged information into the learning process of network, the dependencies and the joint distribution of multiple attributes are exploited to achieve superior aesthetic prediction.

- **A novel method for partially attribute-annotated photos.** We propose a semi-supervised adversarial learning method that can assess the photo aesthetics from partially attribute-annotated photos to reduce the reliance on aesthetic attributes annotation and we are the first to regard the generating processes of score-attributes and the photo as dual tasks [He et al., 2016a], which generates informative feedback signals that benefit both tasks.

- **Summary of domain knowledge in music.** We systematically summarize dimension-related domain knowledge and style-related domain knowledge from musicology and psychology. Then, we annotate two music data sets (MediaEval 2015 and AMG 1608) with musical elements and music styles, which are publicly available at: *https://github.com/GANPerf/Music_PI*

- **A novel learning paradigm in music emotion recognition.** We successfully capture the dependencies between music dimensions and music styles to become new privileged information representation, and further propose the SVR+ model, i.e. learning under new privileged information representation in the standard SVR method of emotion recognition and regression.

- **A novel learning framework for multiple object recognition.** We proposed a novel max-margin learning framework, which can recognize multiple object from images with the help of implicit information about the object's importance conveyed by the list of manually annotated image tags.

- **Extensive experiments.** We conduct extensive experiments on five public data sets to study the behaviour and performance of the proposed framework. Experimental results demonstrate the superiority of the proposed method compared to the state-of-the-art on five benchmark databases.

Chapter 2

Related Work

In this chapter, we first summarize some studies on learning under privileged information paradigm. Then, we discuss its wide applications on different fields.

2.1 Learning Under Privileged Information

Privileged information is the information known as teachers providing students helpful comments, comparisons, and explanations to improve model performance. It is available during training, but it is not available in the testing stage. For example, posterior information (e.g. the post-classification diagnosis of a physician, in a medical setting), future information (e.g. future information about the price of a stock, in a stock-price prediction setting) or hidden information (e.g. hidden information about the importance of objects in multiple object recognition) is informative and can be used as privileged information to enhance learning than using training data alone. The obtaining of privileged information is typically not easily because of higher data collection costs and/or a lack of access to domain expertise. Successfully leveraging privileged information will benefit the learning process of a model.

Vapnik and Vashist [Vapnik and Vashist, 2009] is the first to use privileged information in constructing classifiers. In their model, they are the first to propose learning under privileged information (LUPI) paradigm and they upgrade a support vector machine (SVM) as the SVM+ method where privileged information is used to predict the slack variables in the SVM. Because it is computationally expensive to train SVM+, various research on optimization techniques and SVM+-based algorithms have been proposed to solve SVM+. To solve time-consuming problem on tuning parameters due to L-2 SVM used in LUPI, Niu et al. [Niu and Wu, 2012] propose to use LUPI with L-1 SVM, which is essentially a linear programming to reduce the workload. Lapin et al. [Lapin et al., 2014] propose a paradigm W-SVM where they encode they the prior knowledge as privileged features. They propose a weighted learning process where the prior knowledge is allowed to interpret the

effect of different privileged features. Their model proved to be superior to both the SVM and the SVM+. Li et al. [Li et al., 2016a] propose two efficient algorithms to solve linear SVM+ and kerne SVM+, respectively. For linear SVM+, they propose a new optimization problem which contains simpler constraints. This new optimization problem can be solved by the proposed dual coordinate descent algorithm. For kernel SVM+, they use the ℓ_2- loss to reduce the half of dual variable compared to the original SVM+ method. Li et al. [Li et al., 2018] propose a Robust SVM+ (R-SVM+) algorithm to solve the unclean problem in the training and testing data. In their method, traditional SVM+ method is upgraded as a novel regularization function to cope with the potential noises in the data. The proposed R-SVM+ is efficiently optimized by off-the-shelf solvers.

In addition to SVM+ etc. classification problems, privileged information is also used for regression. Cai et al.[Cai, 2011] combine SVM+ method and multi-tasking learning and form a new methodology called SVM+MTL method for regression. Their method use regression function in decision space and use correcting functions in correcting space. Then proposed model maps data into these two spaces and use the two functions for better prediction. Sarafianos et al. [Sarafianos et al., 2016] propose to use LUPI framework combining human metrology to estimate the height. Due to privileged information unavailable in testing, they train regression models predict the PI in testing, then use the predicted privileged information cooperation with observable features to estimate the height. Shu et al. [Shu et al., 2020] propose a framework called PI-DCNN that use high-level descriptive attributes including photo-related and photography-related attributes as privileged information in the architecture of deep convolutional neural network to predict the aesthetic scores.

2.2 Attribute-Aware Photo Aesthetic Assessment

In this Section, we focus on several works that utilize attributes for photo aesthetic assessment. We divide these works into three categories: low-level attributes, deep features, and high-level attributes. A comprehensive survey related to photo aesthetic assessment can be found in [Yang et al., 2019; Deng et al., 2017; Joshi et al., 2011].

2.2.1 Low-Level Attributes

In early researches, low-level attributes are extensively studied for aesthetic assessment. For example, Lowe [Lowe, 2004] presents an engineering feature approach to transform image data into scale-invariant coordinates relative to local features. These features are shown to provide robust matching across a substantial range of affine

distortion, change in 3D viewpoint, addition of noise, and change in illumination. Ke *et al.*[Ke et al., 2006] propose three distinguishing factors, simplicity, realism, and basic photographic technique, making a photo high-quality or low-quality. Then, the spatial distribution of edges, color distribution, hue count, blur, contrast, and brightness are designed. Luo *et al.* [Luo and Tang, 2008] propose to use the subject region from a photo, then formulated clarity contrast feature, lighting feature, simplicity feature, and composition geometry feature based on the subject and background division. However, low-level features focus more on local hints instead of the global and contextual modeling of a photo. Therefore, they can not totally capture high-level attributes, which is the unavoidable weakness of engineering feature approaches.

2.2.2 Deep Features

With the development and application, deep features, extracted by deep convolutional neural networks, can capture the comprehensive understanding of the input photo. Lu *et al.*[Lu et al., 2014] use a novel double-column deep convolutional neural network to enable automatic feature learning. The proposed deep network unified the feature learning and classifier training, which helps to extract both global and local features of images. Kao *et al.*[Kao et al., 2017] find that semantic recognition task is vital for aesthetic quality prediction. Thus they cast the aesthetic prediction as the main task and captured the relations between semantic recognition task and aesthetic quality prediction via the proposed multi-task deep learning framework. Cui *et al.*[Cui et al., 2018] find that aesthetic perceiving is coupled with semantic understanding. Thus they develop a novel network framework, semantic-aware hybrid network (SANE), to harvest the information from object categorization and scene recognition via this network to improve the performance of aesthetic prediction. For their methods, attributes as input are needed in training. During testing, the attributes are still typically firstly predicted by the model or captured by database. This is usually complex and hardly satisfied in reality.

2.2.3 High-Level Attributes

2.2.3.1 Fully attribute-annotated photos

Several researchers have leveraged aesthetic attributes to facilitate photo aesthetic assessment. For example, Malu *et al.* [Malu et al., 2017] propose a novel multi-task deep convolution neural network (DCNN), which jointly learns high-level attributes along with the overall aesthetic score. In Malu's method, all aesthetic attributes are

needed in their designed multi-task deep convolution neural network (DCNN). They develop a visualization technique to understand the internal representation of these attributes and qualitatively analyze the diverse and complex association with these different attributes. Shu *et al.* [Shu et al., 2021] proposes support vector regression with variational privileged information model and uses high-level descriptive attributes as privileged information to involve the learning process of their proposed method. In their method, all attributes are needed in the training phase, which provides the guidance of predicting the slack variable of support vector regression.

2.2.3.2 Partially attribute-annotated photos

Some studies focus on photo aesthetic from partially attribute-annotated samples. Pan *et al.* [Pan et al., 2019a] propose a multi-task adversarial learning method to simultaneously learn aesthetic attributes and aesthetic score. Their method uses the available attributes as a target to help the network construct better feature representations for aesthetic assessment. Kong *et al.* [Kong et al., 2016] add a branch to predict the aesthetic attributes upon the penultimate layer of the original network. It is helpful only when attribute annotations are available. Then, the final aesthetic score is given based on the features of the aesthetic attributes and content. Although aesthetic attributes can be partially missing in these works, they still require a large amount of aesthetic attributes, and increasing the number of aesthetic attributes is definitely beneficial for photo aesthetic assessment.

2.3 Music Emotion Recognition

In this section, we summarize several works on music emotion recognition methods based on different features that they use during the modeling phase from three aspects, i.e. low-level feature features, deep features and high-level features. A comprehensive survey related to features-based music emotion recognition can be found in [Yang et al., 2018][Panda et al., 2020].

2.3.1 Low-level features

In early research, handcraft acoustic features are extensively studied and used for music emotion recognition. For example, Wang *et al.*[Wang et al., 2012b] propose an acoustic emotion Gaussian (AEG) model to learn a set of latent feature classes from data in order to bridge the acoustic feature space and music emotion space. Deng *et al.*[Deng et al., 2015] apply three kinds of acoustic features: arousal-based features such as RMS, low energy etc.; valence-based features such as zero cross, flux

etc.; and resonance-based features such as roughness, key clarity etc. into regression methods to automatically classify a piece of music. Xianyu *et al.*[Xianyu et al., 2016] propose Double-scale Support Vector Regression (DS-SVR) using 260-dimension low-level handcraft features for music emotion estimation. However, low-level handcraft acoustic features focus more on local hints instead of the global and contextual modelling of songs and they cannot totally capture high-level descriptive information. This is an unavoidable weakness of engineering feature approaches.

2.3.2 Deep features

Deep features, extracted by deep convolutional neural networks, automatically discover the representations for classification or estimation of audios and videos emotion. Mao *et al.*[Mao et al., 2014] propose the use of convolutional neural networks to learn the affect-salient and discriminative features, which leads to stable and robust recognition performance. Choi *et al.*[Choi et al., 2016] use fully convolutional neural networks (FCNs) to achieve automatic music tagging. In their FCN network, the proposed method models the temporal and harmonic structure of audio signals. This makes the features learned from FCNs more beneficial than the handcraft and statistically aggregated features. Han *et al.*[Han et al., 2016] use convolutional neural networks to automatically discover feature representation to recognize the predominant instrument in polyphonic music. Chen*et al.* [Chen et al., 2016] propose deep neural network detector to extract event, object and scene features. These features were integrated via a fusion network to generate a unified representation for video emotion recognition. However, in deep engineering feature approaches, raw input data need to be appropriately preprocessed to generate a good feature representation, which is usually a complex and unavoidable weakness. Moreover, deep features do not thoroughly exploit the domain knowledge of the inherent psychological relationships between human emotion and music.

2.3.3 High-level features

High-level descriptive music features have fallen under the spotlight. Patra *et al.*[Patra et al., 2013] utilize different emotion-related features, i.e. rhythm, timbre and intensity, in the proposed unsupervised classifier for music emotion classification. Panda *et al.*[Panda et al., 2018] capture the MIDI notes from audio signal, melodic features, dynamics features, rhythmic features, musical texture features and expressivity features according to clues like melodic lines, notes, intervals and scores for advancing music emotion recognition. Hu *et al.*[Hu and Yang, 2017] focus on the music style of pop from Western and Chinese songs for music mood prediction via cross

datasets. Shu *et al.*[Shu and Xu, 2019] propose the use of high-level music features such as tempo, mode, brightness etc. as constraints and derive their probabilistic dependencies between high-level music features and emotion for music emotion estimation. Eerola [Eerola, 2011] presents a systematic study of the link between music emotions and music genres. They find that emotions, especially those that express valence, operate differently depending on the musical genre. koutras [Koutras, 2017] proposes a song emotion recognition system to investigate the effect of genre and find that song's genre can significantly improve the performance of music emotion. However, high-level perceptual features and semantic features cannot be regards similarly as engineering feature. A systematical summarization of domain knowledge between high-level descriptive music features and music emotion from musicology and psychology need to be built. Also, a efficient model learning with these high-level descriptive music features should be exploited (i.e. LUPI).

2.4 Multiple Object Recognition

Multiple object recognition from photos has become an important topic in computer vision. A comprehensive survey related to object detection or recognition can be found in [Zhao et al., 2019]

2.4.1 Object Recognition with Traditional Machine Learning Method

In early research, some traditional meachine learning methods are used for multiple object recognition. Felzenszwalb *et al.*[Felzenszwalb et al., 2008] combine a margin-sensitive approach for data mining hard negative examples with a formalism called latent SVM, to allow the effective use of more latent information for object detection tasks. Hwang *et al.*[Hwang and Grauman, 2011] combines SVM and three novel implicit features from an image tag to improve both accuracy and efficiency when detecting the tagged objects. Dvornik *et al.*[Dvornik et al., 2019] uses a convolutional neural network to improve object detection and a CNN network is used to predict the suitability of an image region for placing a given object.

2.4.2 Object Recognition with Deep Learning Method

In current studies, with the development of deep learning techniques, significant advances in the design of network structures and training strategies has attracted much research attention for object recognition in recent years. Schwartz *et al.*[Schwartz et al., 2018] propose a effective framework, which consists of a modified auto-encoder,

denoted Delta-encoder, to synthesize new samples for an unseen category just by seeing few examples. Radovic *et al.*[Radovic et al., 2017] propose to use a convolutional neural networks (CNNs) on a set of aerial images. Their method shows that by selecting a proper set of parameters, a CNN can detect and classify objects with a high level of accuracy.

2.5 Discussion of Privileged Machine Learning

Privileged information has also been used in various fields, such as image categorization [Li et al., 2014][Yan et al., 2016], facial expression recognition [Pan et al., 2019b], domain adaptation [Motiian, 2019][Sarafianos et al., 2017] and deep learning [Lambert et al., 2018][Kamienny et al., 2020]. In this section, we will discuss what kinds of privileged information are available and how the proposed method can be applied in some current multimedia computing tasks.

Discussion We consider examples of multimedia computing tasks that can be used and improved upon by the proposed method. Example 1. It is assumed that in social multimedia networks, the modelling goal is to predict future links in a growing network based on the use of the existing network structure [Qi et al., 2013a]. We have two networks: one is the existing link information in a mature network, and the other is a relatively new network. The derived attributes between node pairs, such as structural similarity or attribute similarity, can be used as privileged information during training to generate predicting links in the new network. Example 2. Let the proposed goal be to group recommendations (e.g. recommending Flicker groups) [Wang et al., 2012a]. Specifically, we must design a system that can index and retrieve groups to help users conveniently search and discover groups of interest. It is beneficial to regard similar latent interests and themes as privileged information between users and recommended groups to allow users access desired groups more easily. Example 3. It is assumed that the proposed goal is to find a typicality ranking scheme for natural scene categorization [Tang et al., 2007b]. In this problem, we consider certain high-level photo attributes such as object emphasis and color because they represent the human perspective. However, in the training data, we have observations about these attributes. Can these be used as privileged information to construct a better prediction rule of typicality ranking? Example 4. Let the proposed goal be automatic semantic annotation of video or video segments [Tang et al., 2007a]. For certain frames, specifically key frames, we can manually annotate the object's importance conveyed in frame tags, which can be used as privileged information during training to improve content-based video searching. PI also can be used in community detection [Qi et al., 2013b].

Chapter 3

Preliminaries

In this chapter, we first introduce and summarize some studies on machine learning algorithms including support vector regression, generative adversarial learning, semi-supervised learning and multi-task learning.

3.1 Support Vector Regression

We first briefly review SVR. The SVR can be considered as one of the most popular regression models. One variant called ϵ-insensitive SVR is given by

$$\begin{aligned} \min_{\theta,\eta^+,\eta^-} \ & \frac{1}{2}||\mathbf{w}||^2 + \gamma \sum_i (\eta_i^+ + \eta_i^-) \\ \text{s.t. } & \mathbf{w}^T \phi(\mathbf{x}_i) + b - y_i \leq \epsilon + \eta_i^+ \\ & y_i - \mathbf{w}^T \phi(\mathbf{x}_i) - b \leq \epsilon + \eta_i^- \\ & \eta_i^+, \eta_i^- \geq 0, \ \forall i \end{aligned} \tag{3.1}$$

where ϵ is a constant that defines the maximum deviation allowed for a prediction to be considered as correct. ϕ is a kernel function that maps $\mathbf{x_i}$ into another feature space. The performance of SVR relies heavily on the choice of kernel function. γ is a trade-off between the model complexity and model loss.

3.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) was first proposed by Goodfellow [Goodfellow et al., 2014] and is now widely applied in many areas such as computer vision and image processing. This is a class of machine learning systems in which two neural networks contest with each other in a zero-sum game framework. The generative network learns to map from a latent space to a data distribution of interest, while the discriminator network distinguishes candidates produced by the generator from the

true data distribution. The generative network's training objective is to increase the error rate of the discriminator network, i.e., "fool" the discriminator network by producing novel candidates that the discriminator identifies as not synthesized (are part of the true data distribution) [Goodfellow et al., 2014]. Specifically, this framework includes two models simultaneously trained: a generative model G that captures the data distribution, and a discriminator model D that estimates the probability that a sample came from the training data. Through an adversarial process, we iteratively evaluate the generative models to determine which one is the best for fitting real data distribution. The objective function for jointly training the prediction model from two networks is:

$$\min_{\theta_g} \max_{\theta_d} [\mathbb{E}_{x \sim Pdata} log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} log(1 - D_{\theta_d}(G_{\theta_g}(z)))], \tag{3.2}$$

where $D_{\theta_d}(x)$ is Discriminator outputting for real data x, $D_{\theta_d}(G_{\theta_g}(z))$ is Discriminator outputting for generated predicted data G(z). In this joint model:

- *Generator network:* Try to fool the discriminator by generating real-looking labels. Specifically, Discriminator (θ_d) wants to maximize the objective such that D(x) is close to 1 (real) and D(G(z)) is close to 0 (prediction).
- *Discriminator network:* Try to distinguish between the real and predicted label. Specifically, Generator (θ_g) wants to minimize objectives such that D(G(z)) is close to 1 (discriminator is fooled into thinking generated G(z) is real).

3.3 Semi-supervised Adversarial Learning

Semi-supervised learning is an approach to machine learning that combines a small amount of labeled data with a large amount of unlabeled data during training. Semi-supervised learning falls between unsupervised learning (with no labeled training data) and supervised learning (with only labeled training data). It is a special instance of weak supervision [Cabannes et al., 2021].

Recent years have seen a few works incorporating adversarial learning with semi-supervised learning. Lai *et al* [Lai et al., 2017] propose a generative adversarial network for learning optical flow in a semi-supervised manner where an adversarial loss was served as guidance for estimating optical flow from both labeled and unlabeled datasets to learn the structural patterns of the flow warp error. Miyato *et al.* [Miyato et al., 2018] propose a virtual adversarial loss function without label information and applied this loss function into their proposed regularization method, which applies to semi-supervised learning. Dong *et al.* [Dong and Lin, 2019] propose a margin

generative adversarial network (MarginGAN) with three players, i.e. a generator, a discriminator and a classifier for semi-supervised learning. Their proposed network generate some pseudo labels, which are used for generated and unlabeled examples in training to make semi-supervised learning meet a variety of practical needs.

3.4 Multi-task Learning

Multitask Learning is an approach to inductive transfer that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias. It does this by learning tasks in parallel while using a shared representation; what is learned for each task can help other tasks be learned better [Caruana, 1997].

In this thesis, we propose a novel framework on privileged machine learning on photo aesthetic assessment topic, where a deep multi-task network as generator is used to learn the aesthetic attributes and score simultaneously to capture their dependencies and extract better feature representations. The deep multi-task method incorporates high-level aesthetic attributes as privileged information (PI), from which better feature representations are constructed via multi-task learning.

Chapter 4

V-SVR+: Support Vector Regression with Variational Privileged Information

4.1 Introduction

Many regression tasks encounter an asymmetric distribution of information between training and testing phases where the additional information available in training, the so-called privileged information (PI), is often inaccessible in testing. In practice, the privileged information in training data might be expressed in different formats, such as continuous, ordinal, or binary values. However, most the existing learning using privileged information (LUPI) paradigms primarily deal with the continuous form of PI, preventing them from managing variational PI, which motivates this research.

In this work, we enhance regression learning with different forms of privileged information. Unlike previous studies, we integrate three forms of privileged information into the regression model to improve regression learning performance. Specifically, the proposed approach with continuous, ordinal, and binary values is based on the maximum margin regression model. For continuous privileged information in the proposed paradigm, we solve this problem using the Lagrangian multiplier method [Everett III, 1963]. For ordinal privileged information in the proposed paradigm, we use the alternating direction method of multipliers (ADMM) due to the large constraints in the constraint condition and its compactness in solving this problem [Boyd et al., 2011]. For binary privileged information, we use stochastic gradient descent (SGD) to solve this optimization problem because the objective function is a differential equation and the partial derivative of the required parameter is easily calculated. The proposed V-SVR+ framework is evaluated in three applica-

tions: photo aesthetic assessment enhanced by high-level aesthetic attributes hidden in the photos; music emotion recognition from songs with the assistance of implicit information about music elements and musical styles judged by composers; multiple object recognition from images with the assistance of implicit information about object importance conveyed by the list of manually annotated image tags.

4.2 Methodology

In this section, we define the problem and introduce the method of SVR+ integrating different forms of privileged information. The framework of the proposed method is shown in Figure 4.1.

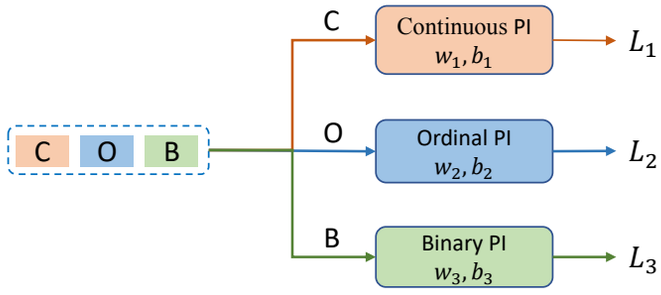


Figure 4.1: The framework of the proposed V-SVR+ model, consisting of three modules to manage three forms of PI respectively. The input of $\{C, O, B\}$ represents continuous, ordinal and binary forms of privileged information. w_1, b_1, w_2, b_2, w_3, b_3 are their parameters.

We denote a set of $V = \{\mathbf{X}, \mathbf{X}^{*(\mathbf{c})}, \mathbf{X}^{*(\mathbf{o})}, \mathbf{X}^{*(\mathbf{b})}, \mathbf{Y}\}$, where $\mathbf{X} \in \mathbb{R}^{N*d}$ is a feature matrix, $\mathbf{Y} \in \mathbb{R}^{N}$ is the ground-truth vector, $\mathbf{X}^{*(\mathbf{c})} \in \mathbb{R}^{N*d^{*(c)}}$, $\mathbf{X}^{*(\mathbf{o})} \in \mathbb{R}^{N*d^{*(o)}}$ and $\mathbf{X}^{*(\mathbf{b})} \in \mathbb{R}^{N*d^{*(b)}}$ are their corresponding continuous, ordinal and binary privileged information matrices. The objective function of the proposed V-SVR+ is given as:

$$\begin{aligned} &\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 + \lambda_3 \mathcal{L}_3 \\ &\text{s.t. } \lambda_1 + \lambda_2 + \lambda_3 = 1 \end{aligned} \tag{4.1}$$

where $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ are the different losses corresponding to the continuous, ordinal and binary forms of PI respectively. λ_1, λ_2 and λ_3 are the weights to balance the tradeoff among losses. The goal of this method is to predict unknown labels for new samples via a real-valued regression function $y = f(x)$.

4.2.1 SVR with Continuous Privileged Information

Given the continuous form of privileged information, we consider the following three linear functions:

$$\begin{aligned} f(\mathbf{X_i}) &= (\mathbf{w}, \phi(\mathbf{X_i})) + b, \\ f_1^*(\mathbf{X_i^{*(c)}}) &= (\mathbf{w_1^*}, \phi(\mathbf{X_i^{*(c)}})) + b_1^*, \\ f_2^*(\mathbf{X_i^{*(c)}}) &= (\mathbf{w_2^*}, \phi(\mathbf{X_i^{*(c)}})) + b_2^* \end{aligned} \tag{4.2}$$

where w, w_1^* and w_2^* are the weight coefficients of the general feature regressor and privileged information regressor, respectively, b, b_1^* and b_2^* are their bias. $\phi(X_i)$ and $\phi(X_i^{*(c)})$ project the i-th training data X_i and $X_i^{*(c)}$ into the kernel space; (,) indicates the matrix product of the two terms in the bracket; and f_1^* and f_2^* are the correcting functions for slacks η^+ and η^- respectively.

Therefore, the objective function of continuous PI is as follows:

$$\begin{aligned} \min_{w, w_1^*, w_2^*} & \sum_{i=1}^{N} \frac{1}{2}(||\mathbf{w_i}||^2 + ||\mathbf{w_{1i}^*}||^2 + ||\mathbf{w_{2i}^*}||^2) + \gamma \sum_{i=1}^{N}[(\mathbf{w_{1i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{1i}] \\ & + \gamma \sum_{i=1}^{N}[(\mathbf{w_{2i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{2i}^*] \\ \text{s.t. } & Y_i - \mathbf{w_i}^T \phi(\mathbf{X}_i) - b \leq \epsilon + (\mathbf{w_{1i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{1i}^* \\ & \mathbf{w_i}^T \phi(\mathbf{X}_i) + b - Y_i \leq \epsilon + (\mathbf{w_{2i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{2i}^* \\ & [(\mathbf{w_{1i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{1i}^*] \geq 0 \\ & [(\mathbf{w_{2i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{2i}^*] \geq 0, \ \forall i \end{aligned} \tag{4.3}$$

To solve this convex optimization problem, we apply the Lagrangian multiplier method [Everett III, 1963] to simplify the formulation and improve computational efficiency. The Lagrangian function is shown as follows:

$$\begin{aligned} \mathcal{L}_1 = & \sum_{i=1}^{N} \frac{1}{2}(||\mathbf{w_i}||^2 + ||\mathbf{w_{1i}^*}||^2 + ||\mathbf{w_{2i}^*}||^2) \\ & + \gamma \sum_{i=1}^{N}[(\mathbf{w_{1i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{1i}^*] + \gamma \sum_{i=1}^{N}[(\mathbf{w_{2i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{2i}^*] \\ & + \sum_{i=1}^{N} \alpha_i (Y_i - f(\mathbf{X_i}) - \epsilon - f_1^*(\mathbf{X_i^{*(c)}})) + \sum_{i=1}^{N} \beta_i (0 - f_1^*(\mathbf{X_i^{*(c)}})) \\ & + \sum_{i=1}^{N} \alpha_i^* (f(\mathbf{X_i}) - Y_i - \epsilon - f_2^*(\mathbf{X_i^{*(c)}})) + \sum_{i=1}^{N} \beta_i^* (0 - f_2^*(\mathbf{X_i^{*(c)}})) \end{aligned} \tag{4.4}$$

where α, α^*, β, β^* are Lagrangian multipliers.

Then, we obtain the dual problem for this object as follows:

$$\begin{aligned}
&\min_{\alpha,\alpha^*,\beta,\beta^*} \sum_{i=1}^{N} \gamma_i(\alpha_i - \alpha_i^*) + \epsilon(\alpha_i + \alpha_i^*) + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\mathbf{X_i}, \mathbf{X_j}) \qquad (4.5)\\
&+ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^* + \beta_i^* - \gamma)(\alpha_j^* + \beta_j^* - \gamma)K^*(\mathbf{X_i}^{*(c)}, \mathbf{X_j}^{*(c)})\\
&+ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i + \beta_i - \gamma)(\alpha_j + \beta_j - \gamma)K^*(\mathbf{X_i}^{*(c)}, \mathbf{X_j}^{*(c)})\\
&\text{s.t.} \sum_{i=1}^{N}(\alpha_i^* - \alpha_i) = 0, \sum_{i=1}^{N}(\alpha_i^* + \beta_i^* - \gamma) = 0\\
&\sum_{i=1}^{N}(\alpha_i + \beta_i - \gamma) = 0\\
&\alpha_i^*, \alpha_i, \beta_i^*, \beta_i \geq 0 \; i = 1, ..., N.
\end{aligned}$$

where $K(\cdot,\cdot)$ and $K^*(\cdot,\cdot)$ are the kernels that define the inner products for $\phi(\mathbf{X})$ and $\phi(\mathbf{X}^{*(c)})$ respectively; and $\alpha, \alpha^*, \beta, \beta^*$ are the model parameters for solving this optimization problem.

Then, based on the KKT condition, we obtain the parameters w and b using the details in [Platt, 1998] as follows:

$$\begin{aligned}
w &= \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)\mathbf{X}_i,\\
b &= \frac{1}{|S| + |S'|}\left[\sum_{s\in S}(\mathbf{Y_s} - \epsilon - w\mathbf{X_s}) + \sum_{s\in S'}(\mathbf{Y_s} + \epsilon - w\mathbf{X_s})\right],
\end{aligned} \qquad (4.6)$$

where S and $S^{'}$ are sets that correspond to two correcting functions The two correcting functions are shown as follows:

$$\begin{aligned}
f_1^*(\mathbf{X}) &= (w_1^*, \phi(\mathbf{X}^{*(c)})) + b_1^* = \sum_{i=1}^{N}(\alpha_i + \beta_i - \gamma)K^*(\mathbf{X_i}^{*(c)}, \mathbf{X}^{*(c)}) + b_1^*,\\
f_2^*(\mathbf{X}) &= (w_2^*, \phi(\mathbf{X}^{*(c)})) + b_2^* = \sum_{i=1}^{N}(\alpha_i^* + \beta_i^* - \gamma)K^*(\mathbf{X_i}^{*(c)}, \mathbf{X}^{*(c)}) + b_2^*
\end{aligned} \qquad (4.7)$$

Thus, the decision function for solving this problem is given as follows:

$$f(\mathbf{X}) = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)K(\mathbf{X_i}, \mathbf{X}) + b. \qquad (4.8)$$

We use the conditional gradient method to solve Eq. 6.17. The obtained parameters

w and b as w_1 and b_1 in the first module of the proposed framework as shown in Figure 4.1.

4.2.2 SVR with Ordinal Privileged Information

For ordinal privileged information, we consider two forms of ranking for different tasks: monotonically ascending ranking and monotonically descending ranking. Thus, we provide the following ranking sets of privileged information: $\mathbf{E} = \{(i,j)|\mathbf{X}_i^{*(o)} \leq \mathbf{X}_j^{*(o)} \text{ or } \mathbf{X}_i^{*(o)} > \mathbf{X}_j^{*(o)}\}$, where $i,j \in \{1,2...,d^{*(o)}\}$. Specifically, (i,j) when $\mathbf{X}_i^{*(o)} \leq \mathbf{X}_j^{*(o)}$ denotes an ascending relationship among PI. Conversely, when $\mathbf{X}_i^{*(o)} > \mathbf{X}_j^{*(o)}$, a descending relationship is indicated.

In this situation, two mappings can be represented as follows:

$$\begin{aligned} f(\mathbf{X}_i) &= (\mathbf{w}, \phi(\mathbf{X}_i)) + b, \\ f^*(\mathbf{X}_i^{*(o)}) &= (\mathbf{w}^*, \phi(\mathbf{X}_i^{*(o)})) + b^*, \end{aligned} \tag{4.9}$$

where w^* and b^* are the weight coefficients and bias of the privileged information respectively; (,) indicates the matrix product of two terms in brackets. Then, we have the following constraints:

$$\begin{aligned} |Y_i - f(\mathbf{X}_i)| &\leq \epsilon + \eta_i, \\ |Y_i - f^*(\mathbf{X}_i^{*(o)})| &\leq \epsilon + \eta_i, \end{aligned} \tag{4.10}$$

where η_i is the slack variable that measures the failure to meet ϵ distance.

Then, we consider the defined set $\mathbf{E}$ and provide the following inequation:

$$f^*(\mathbf{X^{*(o)}}_i) - f^*(\mathbf{X^{*(o)}}_j) \geq 1 - \xi_{ij}^*, \tag{4.11}$$

where ξ_{ij}^* is the slack variable to allow some number of disorders where the important privileged information is ranked below the unimportant privileged information, and i and j satisfy ranking sets $\mathbf{E}$.

Thus, the objective function handling with ordinal PI is as follows:

$$\begin{aligned}
&\min_{\theta,\eta,\xi} \frac{1}{2}(||\mathbf{w}||^2 + ||\mathbf{w}^*||^2) + \gamma_1 \sum_{n=1}^{N}\sum_{k=1}^{N}(\eta_k^{(n)+} + \eta_k^{(n)-}) + \gamma_2 \sum_{n=1}^{N}\sum_{(i,j)\in\mathbf{E}} \xi_{ij}^{*(n)} \\
&\text{s.t. } \mathbf{w}^T\mathbf{X}_k^{(n)} + \mathbf{w}^{*T}\mathbf{X}^{*(\mathbf{o})}{}_k^{(n)} + b + b^* - Y_k^{(n)} \leq \epsilon + \eta_k^{(n)+} \\
&\quad Y_k^{(n)} - \mathbf{w}^T\mathbf{X}_k^{(n)} - \mathbf{w}^{*T}\mathbf{X}^{*(\mathbf{o})}{}_k^{(n)} - b - b^* \leq \epsilon + \eta_k^{(n)-} \\
&\quad \mathbf{w}^{*T}(\mathbf{X}^{*(\mathbf{o})}{}_i^{(n)} - \mathbf{X}^{*(\mathbf{o})}{}_j^{(n)}) \geq 1 - \alpha_{ij}\xi_{ij}^{*(n)} \\
&\quad \eta_k^{(n)+}, \eta_k^{(n)-}, \xi_{ij}^{*(n)} \geq 0 \\
&\quad (i,j) \in \mathbf{E}, n = 1, ..., N
\end{aligned} \tag{4.12}$$

where $\theta = [\mathbf{w}; \mathbf{w}^*; b; b^*]$ are the parameters to be optimized. The first four sets of constraints are used to fit the regression model to the ground-truth labels. The fifth set of constraints is used to exploit the ordinal privileged information. The additional parameter $\alpha = \{\alpha_{ij}\} = \{\frac{1}{|i-j|}\}, \forall i \neq j$ is used to satisfy temporal smoothness.

We use ADMM when the optimization problem contains a large number of constraints. The usage of augmented Lagrangian multipliers can accelerate convergence [Boyd et al., 2011]. The complexity for solving Eq (4.12) is $O(n^2)$, where n is the number of samples.

Because ADMM cannot be directly applied to Eq. (4.12), we first transform Eq. (4.12) and define $\mu = [\gamma_1; \gamma_2\alpha]$, where γ_1 is the coefficient of the first four sets of constraints, $\gamma_2\alpha$ is the fifth set of constraints in Eq. (4.12) and $\mu \in \mathbb{R}^{M_1+M_2}$ is a vector whose first M_1 entries are γ_1 and the last M_2 entries are $\gamma_2\alpha$. Let $t = [\eta_k^{(n)+}, \eta_k^{(n)-}, \xi_{ij}^{*(n)}] \in R^{M_1+M_2}$, where M_1 is the number of constraints corresponding to $[\eta_k^{(n)+}, \eta_k^{(n)-}]$ loss and M_2 is the number of constraints corresponding to $\xi_{ij}^{*(n)}$ loss. Based on these definitions, we formulate Eq. (4.12) as

$$\begin{aligned}
&\min_{\theta,t} \frac{1}{2}\theta^T\Lambda\theta + \mu^T t \\
&\text{s.t. } A\theta - t = c
\end{aligned} \tag{4.13}$$

where $\Lambda \in \mathbb{R}^{(d+d^{*(o)}+2)\times(d+d^{*(o)}+2)}$ is a diagonal matrix and $A \in \mathbb{R}^{(M_1+M_2)\times(d+d^{*(o)}+2)}$ is a matrix and $c \in \mathbb{R}^{M_1+M_2}$. Specifically,

$$A = \begin{bmatrix} \mathbf{X} & \mathbf{X}^{*(\mathbf{o})} & \mathbf{1} & \mathbf{1} \\ -\mathbf{X} & -\mathbf{X}^{*(\mathbf{o})} & -\mathbf{1} & -\mathbf{1} \\ \mathbf{0} & -\mathbf{X}_E & \mathbf{0} & \mathbf{0} \end{bmatrix} c = \begin{bmatrix} \epsilon\mathbf{1} + \mathbf{Y} \\ \epsilon\mathbf{1} - \mathbf{Y} \\ -\mathbf{1} \end{bmatrix} \tag{4.14}$$

where $\mathbf{X}_E$ is a matrix whose rows are the difference between two PIs whose indices

belong to set **E**. **1** and **0** are vectors with the proper dimensions containing all 1s and 0s, respectively. **Y** is a vector of the known labels.

The augmented Lagrangian has a quadratic form with respect to θ, t and is linear to v. The augmented Lagrangian can be formulated as the following equation which can be solved by ADMM.

$$\mathcal{L}_2 = L_\rho(\theta, t, v) = \frac{1}{2}\theta^T \Lambda \theta + \mu^T t + v^T(A\theta - t - c) + \frac{\rho}{2}||A\theta - t - c||_2^2 \tag{4.15}$$

Then, the gradient of $L_p(\theta, t, v)$ w.r.t. θ or t can be computed as follows, respectively.

$$\frac{\partial L_\rho(\theta, t, v)}{\partial \theta} = \theta\Lambda + v^T A + \rho A^T ||A\theta - t - c||_1 \tag{4.16}$$

$$\frac{\partial L_\rho(\theta, t, v)}{\partial t} = \mu^T - v^T - \rho A||A\theta - t - c||_1 \tag{4.17}$$

The new iterations of θ, t and v can be produced by letting $\partial L_p(\theta, t, v)/\partial\theta = 0$ and $\partial L_p(\theta, t, v)/\partial t = 0$.

$$\theta^{k+1} := [\frac{1}{\rho}\Lambda + A^T A]^{-1} A^T (t^k - \frac{1}{\rho}v^k + c) \tag{4.18}$$

$$t_i^{k+1} := \frac{1}{\rho}v^k + A\theta^{k+1} - \frac{1}{\rho}\mu - c \tag{4.19}$$

$$v^{k+1} := v^k + \rho(A\theta^{k+1} - t^{k+1} - c) \tag{4.20}$$

Let $f(\theta) = \theta^T \Lambda \theta$ and $g(t) = \mu^T t$. Because $f(\theta)$ and $g(t)$ are convex functions, L_ρ in Eq. 7.9 is also convex. Thus, Eq. 4.18 and Eq. 4.19 can obtain the unique optimal solution and these updates are convergent [Boyd et al., 2011]. In particular, $t = [\eta^{(n)+}, \eta^{(n)-}, \xi^{*(n)}]$ is a non-negative, and we set a threshold λ to make sure $t \geq 0$. Let $z_i = \frac{1}{\rho}v^k + A\theta^{k+1} - \frac{1}{\rho}\mu - c$, then

$$\mathbf{t}_i^{k+1} = T_\lambda(z_i) = \begin{cases} z_i - \lambda, & \text{if} \quad z_i > \lambda \\ 0, & \text{if} \quad |z_i| \leq \lambda \\ z_i + \lambda, & \text{if} \quad z_i < -\lambda \end{cases} \tag{4.21}$$

where $\lambda \geq 0$ is the soft threshold operator and i is the ith entry in each vector.

After finishing the optimization problem, the estimated parameters w and b as w_2 and b_2 in the second module of the proposed framework are shown in Figure 4.1.

4.2.3 SVR with Binary Privileged Information

In this section, privileged information $X_{im}^{*(b)} \in \{0,1\}_{m=1}^{d^{*(b)}}$ is a binary or Boolean value that denotes whether the i-th training data contains privileged information. Hypothesis $Y_i \in [a,b]$ and q is a threshold to divide Y_i into good performance $[a,q]$ as S^+ and bad performance $(q,b]$ as S^-. In general, because privileged information provides the model with helpful comments, comparisons, and explanations, we conclude that the model with PI is more likely to improve performance. Conversely, if PI is not involved in the learning process of the model, the performance will worsen. Thus, the probabilistic relations are shown as:

$$\begin{aligned} p(\hat{Y} = S^+ | X^{*(b)} = 1) > p(\hat{Y} = S^- | X^{*(b)} = 1) \\ p(\hat{Y} = S^- | X^{*(b)} = 0) > p(\hat{Y} = S^+ | X^{*(b)} = 0) \end{aligned} \tag{4.22}$$

where $p(\hat{Y} = S^+ | X^{*(b)} = 1)$ and $p(\hat{Y} = S^- | X^{*(b)} = 1)$ indicate the probabilities of good performance and poor performance respectively, when there is helpful privileged information. $p(\hat{Y} = S^- | X^{*(b)} = 0)$ and $p(\hat{Y} = S^+ | X^{*(b)} = 0)$ are the complementary expressions.

In the proposed method, we use the ReLU function to penalize samples that violate this formula. The corresponding penalty $l_{im}(\mathbf{X_i}, X_{im}^{*(b)}, \hat{Y}_i)$ based on Eq. 4.22 is encoded as follows:

$$\begin{aligned} &\ell_{im}(\mathbf{X_i}, X_{im}^{*(b)}, \hat{Y}_i) = X_{im}^{*(b)} * [p(\hat{Y}_i = S^- | X_{im}^{*(b)} = 1) - p(\hat{Y}_i = S^+ | X_{im}^{*(b)} = 1)]_+ \\ &+ (1 - X_{im}^{*(b)}) * [p(\hat{Y}_i = S^+ | X_{im}^{*(b)} = 0) - p(\hat{Y}_i = S^- | X_{im}^{*(b)} = 0)]_+ \\ &= X_{im}^{*(b)} * [1 - 2 * p(\hat{Y}_i = S^+ | X_{im}^{*(b)} = 1)]_+ + (1 - X_{im}^{*(b)}) * [2 * p(\hat{Y}_i = S^+ | X_{im}^{*(b)} = 0) - 1]_+ \end{aligned} \tag{4.23}$$

where $[\cdot] = max(\cdot, 0)$.

The objective function of binary PI is as follows:

$$\mathcal{L}_3 = \frac{1}{2}\mathbf{w}^T\mathbf{w} + c_1 \sum_{i=1}^{N} \ell_\epsilon(f(\mathbf{X_i}, \mathbf{w}) - Y_i) + c_2 \sum_{i=1}^{N} \sum_{m=1}^{d^{*(b)}} \ell_{im}(\mathbf{X_i}, X_{im}^{*(b)}, \hat{Y}_i) \tag{4.24}$$

where $\mathbf{w}$ is the parameter of the regression model; and c_1 and c_2 are the coefficients. For $f(\mathbf{X}, \mathbf{w})$, we use a linear function and apply the sigmoid function to replace the probabilistic dependencies as follows:

$$\begin{aligned} p(\hat{Y} = S^+ | X^{*(b)}) = \sigma(f(X, w)) \\ p(\hat{Y} = S^- | X^{*(b)}) = 1 - \sigma(f(X, w)) \end{aligned} \tag{4.25}$$

where $\sigma(X) = \frac{1}{1+e^{-X}}$.

We use stochastic gradient descent (SGD) to solve the problem. The updating rule is shown as follows:

$$w^{(t+1)} = w^{(t)} - \eta^{(t)} \frac{\partial \mathcal{L}_3}{\partial w} \tag{4.26}$$

where t and η are the number of iterations and the learning rate respectively.

The gradient of the loss function to the weight can be computed as follows:

$$\frac{\partial \mathcal{L}_3}{\partial \mathbf{w}} = \mathbf{w} + c_1 \sum_{i=1}^{N} \frac{\partial \ell_\epsilon(f(\mathbf{X_i}, \mathbf{w}) - Y_i)}{\partial \mathbf{w}} + c_2 \sum_{i=1}^{N} \sum_{m=1}^{d^{*(b)}} \frac{\partial \ell_{im}(\mathbf{X_i}, X_{im}^{*(b)}, \hat{Y}_i)}{\partial \mathbf{w}} \tag{4.27}$$

where the specific gradient of the loss function to the weight is computed as:

$$\frac{\partial \ell_\epsilon(f(\mathbf{X_i}, \mathbf{w}) - Y_i)}{\partial \mathbf{w}} = \begin{cases} 0, & if \ \ |f(\mathbf{X_i}) - Y_i| \leq \epsilon \\ \phi(\mathbf{X_i}), & otherwise. \end{cases} \tag{4.28}$$

$$\frac{\partial \ell_{im}(\mathbf{X_i}, X_{im}^{*(b)}, \hat{Y}_i)}{\partial \mathbf{w}} = \begin{cases} -2\sigma(f(\mathbf{X_i}, \mathbf{w}))[1 - \sigma(f(\mathbf{X_i}, \mathbf{w}))]\phi(\mathbf{X_i}), \\ if \ X_{im}^{*(b)} = 1 \ and \ 1 - 2\sigma(f(\mathbf{X_i}, \mathbf{w})) \geq 0 \\ 2\sigma(f(\mathbf{X_i}, \mathbf{w}))[1 - \sigma(f(\mathbf{X_i}, \mathbf{w}))]\phi(\mathbf{X_i}), \\ if \ X_{im}^{*(b)} = 0 \ and \ 2\sigma(f(\mathbf{X_i}, \mathbf{w})) - 1 \geq 0 \\ 0, \qquad otherwise. \end{cases} \tag{4.29}$$

When the algorithm converges, the estimated parameters w and b as w_3 and b_3 in the third module of the proposed framework shown in Figure 4.1.

4.3 Conclusion

In this work, we propose a unified framework called V-SVR+ that involves three forms of privileged information: continuous, ordinal, and binary. We design different loss functions and optimization algorithms specific to different forms of PI. We will use three different tasks, photo aesthetic assessment, music emotion recognition and multiple object recognition in the next three chapters, to demonstrate the proposed methods. For the photo aesthetic assessment, we consider whether a photo contains high-level aesthetic attributes as continuous/ordinal/binary privileged information in the proposed model. In music emotion recognition, musical elements such as tempo and brightness are used as continuous privileged information and are integrated into the objective functions. For multiple object recognition, the implicit information about object importance is considered to be ordinal privileged information to enhance the recognition task. We conducted extensive experiments to demonstrate the superiority of the proposed V-SVR+ framework compared to several other

methods with four benchmark databases.

Chapter 5

Learning with Privileged Information for Photo Aesthetic Assessment

5.1 Introduction

Photo aesthetic assessment task aims to endow computers with the ability of perceiving aesthetics as human beings. The objective of this task is to automatically evaluate how beautiful this photo is from an aesthetic perspective. Such assessments have many attractive applications, for example, personal photo album management, automatic photo editing, and photo retrieval. As a result, photo aesthetic assessment has attracted more and more attention in recent years [Kong et al., 2016][Lv et al., 2018][Hosu et al., 2019]. Although much progress has been achieved at building a photo rating classifier, it is a still very challenging subjective task due to heavily relying on human perception to the photos.

To address this issue, early works on photo aesthetic assessment either focus on hand-crafted features, deep features [Cui et al., 2018] or classifiers. Hand-crafted features [Luo and Tang, 2008], such as lighting, contrast, and global image layout are used to represent the photographic rules, and deep features [Lv et al., 2018] are used to represent photo content during feature extraction phase. Then, photo aesthetic assessment task is regarded as a classification or regression problem with support by support vector machine/regression, naive Bayes classifier, deep classification/regression model etc.

Despite of achieving the satisfactory assessment, most of the above approaches do not take into consideration of the high-level attributes, which can explicitly present some photo quality cues that a human expert might use to evaluate a photo aesthetic quality. Figure 7.1 shows examples of how high-level attributes such as color, content, light etc. can affect aesthetic qualities. The first and third rows include photos

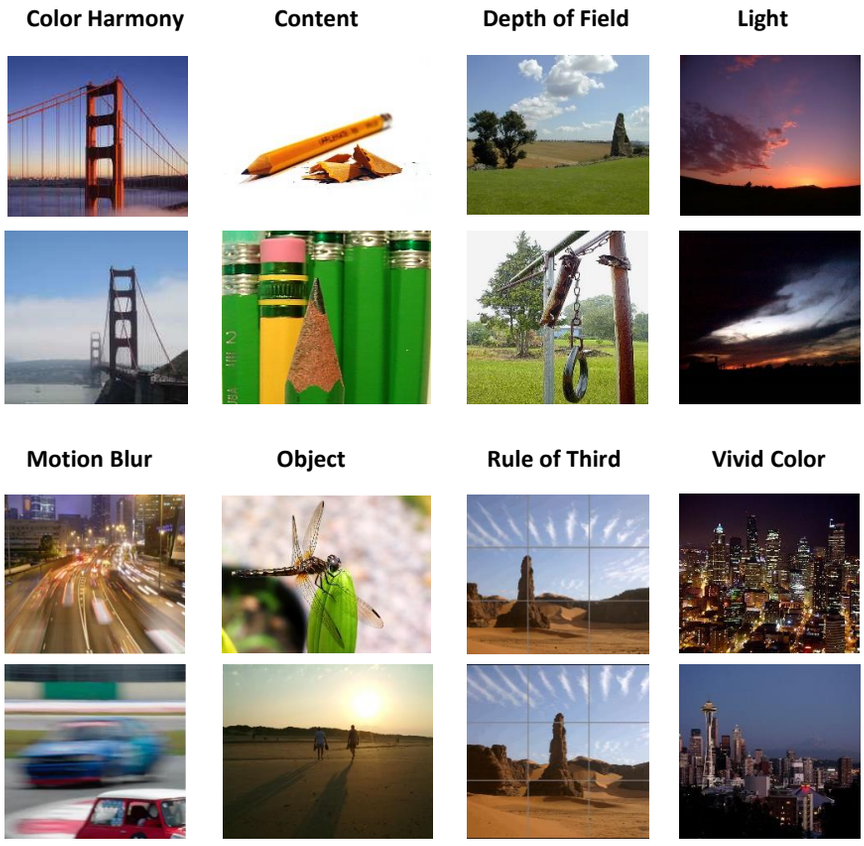


Figure 5.1: Different aesthetic qualities w.r.t. different photo attributes as privileged information. The first and third rows include photos with high aesthetic qualities. The second and fourth rows include photos with low aesthetic qualities.

with high aesthetic qualities, whereas the second and fourth rows indicate the low aesthetic quality photos. In this example, high-level aesthetic photos are usually with harmonious color, abundant content, good light, vivid color, etc, make them fascinating. In the contrast, high-level attributes of disharmonious color, boring content, poor lighting and dull color, etc. determine the low aesthetic photos. Therefore, the above observations conclude that such high-level attributes can largely impact the aesthetics results.

Current works on high-level descriptive photo attributes have several limitations. First, some approaches only utilize one attribute, overlooking the broad relation spectrum between aesthetics and multi-attributes [Cui et al., 2018][Nishiyama et al., 2011][Kao et al., 2017]. Second, multi-attributes deep framework in current works ignore the inherent dependencies among attributes [Malu et al., 2017][Pan et al., 2019a], such as the coexistent and exclusive relations, or ignore the joint distributions among aesthetic attributes [Lv et al., 2018][Shu et al., 2020], which are crucial for photo

aesthetic assessment.

Therefore, in this chapter, we propose a novel method, which utilizes the high-level aesthetic attributes as privileged information to combine reconstructing descriptive attributes and learning the aesthetic score in a simultaneous manner. Specifically, during training, privileged information is used in three aspects: first, we propose to rank photo aesthetics in which the relative ranking of PI is directly modeled in the loss function; second, prior probabilities loss is adopted to capture the coexistent and exclusive relations on privileged information space; third, the adversarial learning is applied in privileged information space to enforce the distribution of the predicted attributes as close as possible to that of ground truth. During testing, the aesthetic score of an unknown photo is predicted by the learned network and the predicted aesthetic attributes can also be given if necessary.

5.2 Preliminaries

Denote a set of triples $S = \{\mathbf{X}, \mathbf{X}^*, \mathbf{Y}\}$, where $\mathbf{X} \in \mathbb{R}^{N*d}$ is a feature matrix and $\mathbf{X}^* \in \mathbb{R}^{N*d^*}$ is a privileged information matrix. Each PI can be either continuous ($X^*_{ik} \in \mathbb{R}$) or binary ($X^*_{ik} \in \{0, 1\}$). (i, k) is the matrix subscript where i represents instance index and k represents PI index. N is the number of training instances. d and d^* are the dimensions of feature and PI respectively. $\mathbf{Y} \in \mathbb{R}^N$ is the ground-truth vector. We want to learn a network which can capture the correlation between the aesthetic score and PI. The goal of regression learning is to learn function $f : \mathbb{R}^d \rightarrow \mathbb{R}^{d^*+1}$ outputting PI and the predicted aesthetic score simultaneously, which is a mapping from the feature space $\mathcal{X}$ to a label space $\mathcal{Y}$ and PI space $\mathcal{X}^*$.

5.3 Methodology

5.3.1 Full Attribute-annotated Photos: Privileged Multi-task Learning for Attribute-aware Aesthetic Assessment

The framework of the proposed method is summarized in Figure 5.2. We adopt a strategy of learning with privileged information, in which aesthetic attributes are treated as privileged information and assist the learning process of the aesthetic assessment network during training phase. During testing phase, the predictions of testing photos are given by the learned aesthetic assessment network. In the following subsections, we first introduce the basic networks. Then, we arrange the guidance of PI in label space by ranking constraint. At the same time, attributes are learned by similarity constraint, prior probabilities loss and adversarial learning in privileged

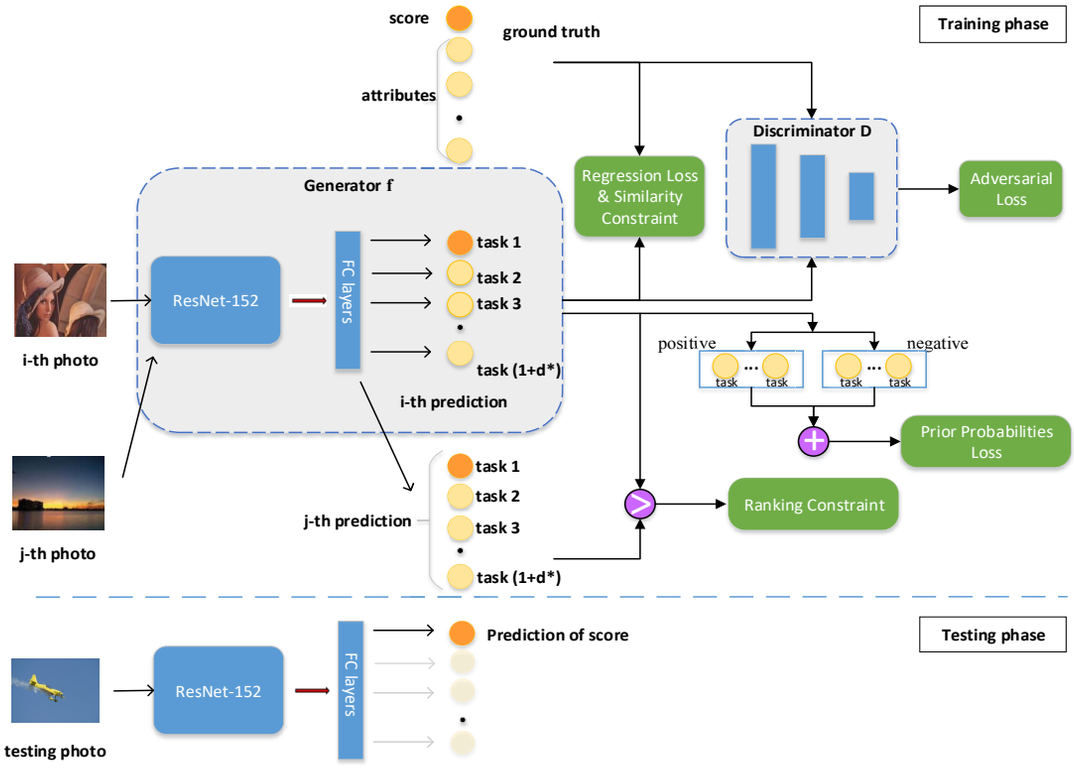


Figure 5.2: The framework of the proposed method. In training phase, two photos are first trained respectively. Utilizing pairwise photos aims to learn ranking loss functions. In our generator networks, $d^* + 1$ tasks correspond to d^* learning tasks for aesthetic attributes and one task for predicting the overall score. These aesthetic attributes enable to provide more detailed feedback on automatic aesthetic assessment. "positive" and "negative" refer to coexistent and exclusive relations among multiple tasks.

information space. Finally, we give the overall loss function and optimization.

5.3.1.1 Basic Network

We first build the deep convolutional neural networks with two photos. The first one $\hat{Y}_i = f(\mathbf{X_i}; \theta)$ performs aesthetic assessment from i-th photo, and is discussed mainly in the next steps. The second one $\hat{Y}_j = f(\mathbf{X_j}; \theta)$ with j-th photo is used for comparison. To make the predicted label $\hat{\mathbf{Y}}$ be approximate to the real label $\mathbf{Y}$, we use regression loss defined by $\ell_{reg} = \sum_{i=1}^{N} ||\hat{Y}_i - Y_i||_2^2$ and $\ell_{reg} = \sum_{j=1}^{N} ||\hat{Y}_j - Y_j||_2^2$, where Y_i and Y_j are the ground-truth rating for photo-i and photo-j, and $\hat{Y}_i$ and $\hat{Y}_j$ are the estimated scores by the CNN model. Regression loss contributes to the capacity of the network on predicting the commonly occurring range of scores.

5.3.1.2 Guidance in Label Space

Ranking Constraint

In this section, we propose the ranking constraint to describe relative ranking of aesthetic attributes and scores in label space. The ranking loss is shown in Eq.5.1.

$$\begin{aligned}\ell_{rank} = &\sum_{k=1}^{d^*}\sum_{i=1}^{N}\sum_{j=1}^{N}\{(X_{ik}^* - X_{jk}^*) - C_k(\hat{Y}_i - \hat{Y}_j)\}^2 \\ &+ \sum_{i=1}^{N}\sum_{j=i+1}^{N}[\beta - T(Y_iY_j)(\hat{Y}_i - \hat{Y}_j)]_+,\end{aligned} \tag{5.1}$$

where $C_k > 0$ is the weight between privileged information and predicted scores, $T(Y_iY_j) = \begin{cases} 1, & if \quad Y_iY_j \\ -1, & if \quad Y_i < Y_j \end{cases}$, $[\cdot]_+ = max(\cdot, 0)$ and β is a specified margin parameter. The first term makes the readily available ordinal information and satisfies the assumption of aesthetic attributes highly related to photo aesthetic assessment. i.e., photo with positive attributes is more possible to have a higher aesthetic score [Dhar et al., 2011]. The second term describes relative aesthetic scores rankings of photo pairs. i.e., if the ground truth score of photo-i is higher than the score of photo-j, the corresponding estimated score of photo-i should be still higher than photo-j. This equation penalizes mistakes for extreme scores more heavily of the network.

5.3.1.3 Guidance in Privileged Information Space

Aesthetic attributes as privileged information in our work are used in two aspects. First, in our designed network, we use extra units upon the output layer to learn privileged information, leading a deep multi-task network. Through the network learning with aesthetic score and privileged information simultaneously, privileged information are explored to construct better feature representations. Second, we propose an adversarial learning to make deep multi-task network and discriminator compete with each other in order to enforce the distribution of the predicted privileged information to close to the ground truth distribution.

Similarity Constraint

The similarity constraint focuses on the network on predicting privileged information upon the output layer closing to ground truth. Since the value of privileged information is continuous or binary, we give two formulations of squared error and

binary cross entropy shown in Eq.5.2 to handle with corresponding forms.

$$\ell_{sim} = \begin{cases} \sum_{i=1}^{N}\sum_{k=1}^{d^*}(X_{ik}^* - \hat{X}_{ik}^*)^2 & if \;\; X_{ik}^* \in \mathbb{R} \\ \sum_{i=1}^{N}\sum_{k=1}^{d^*} -X_{ik}^* \log \hat{X}_{ik}^* - (1 - X_{ik}^*)\log(1 - \hat{X}_{ik}^*) & \\ & if \;\; X_{ik}^* \in \{0,1\}. \end{cases} \tag{5.2}$$

In Eq.5.2, $\hat{X}_{ik}^*$ denotes the numeric of the predicted privileged information, X_{ik}^* denotes binary {0,1} ground truth of sign privileged information, or continuous ground truth of PI. When X_{ik}^* is binary, this formula just calculates $-X_{ik}^* \log \hat{X}_{ik}^*$ or $-(1 - X_{ik}^*)\log(1 - \hat{X}_{ik}^*)$ with X_{ik}^* being 1 or 0 respectively. When X_{ik}^* equals 1, the predicted value $\hat{X}_{ik}^*$ is expected to close to 1 in order to minimize ℓ_{sim} loss function. Similarly, when X_{ik}^* equals 0, the predicted value $\hat{X}_{ik}^*$ is expected to close to 0.

Prior Probabilities Loss

In this section, we exploit the prior probabilistic correlations among attributes. Attributes often appear simultaneously because of inherent relations. We denote this as positive correlations when these dependencies show the co-occurrence of attributes. On the other hand, when some attributes exclude from others, we denote them as negative correlations.

Let $U = U_P \cup U_N$, where U_P denotes the set of positively correlated attribute pairs and U_N denotes the set of negatively correlated attribute pairs. In the set of positively correlated attribute pairs, the probabilistic relations are shown as:

$$p(X_k^* = 1 | X_h^* = 1) > p(X_k^* = 0 | X_h^* = 1) \tag{5.3}$$

$$p(X_k^* = 1 | X_h^* = 1) > p(X_k^* = 1 | X_h^* = 0), \tag{5.4}$$

where $p(X_k^* = 1|X_h^* = 1)$ and $p(X_k^* = 0|X_h^* = 1)$ indicate that the probabilities of k-th appears or not, when h-th attribute appears. $p(X_k^* = 1|X_h^* = 1)$ and $p(X_k^* = 1|X_h^* = 0)$ are the complementary expression. From Eq. 5.3, we get $\frac{p(X_k^*=1, X_h^*=1)}{p(X_h^*=1)} > \frac{p(X_k^*=0, X_h^*=1)}{p(X_h^*=1)}$. From Eq. 5.4, we get $\frac{p(X_k^*=1, X_h^*=1)}{p(X_h^*=1)} > \frac{p(X_h^*=0|X_k^*=1)p(X_k^*=1)}{p(X_h^*=0)}$, then $\frac{p(X_h^*=1|X_k^*=1)p(X_k^*=1)}{p(X_h^*=1)} > \frac{[1-p(X_h^*=1|X_k^*=1)]p(X_k^*=1)}{1-p(X_h^*=1)}$, then we get $p(X_h^* = 1|X_k^* = 1) > p(X_h^* = 1)$. Thus, the equivalent formulas we can get are

$$p(X_k^* = 1, X_h^* = 1) > p(X_k^* = 0, X_h^* = 1) \tag{5.5}$$

$$p(X_k^* = 1, X_h^* = 1) > p(X_k^* = 1)p(X_h^* = 1), \tag{5.6}$$

Similarly, for the set of negatively correlated attribute pairs, we can get the following

formula:

$$p(X_k^* = 1, X_h^* = 1) < p(X_k^* = 0, X_h^* = 1) \tag{5.7}$$

$$p(X_k^* = 1, X_h^* = 1) < p(X_k^* = 1)p(X_h^* = 1), \tag{5.8}$$

Then we convert the estimated attributes to discrete values of 0 or 1 via setting a simply threshold and define $p^{\hat{X}_{k1}^*}$ represents the estimated marginal probability $p(\hat{X}_k^* = 1)$, $p^{\hat{X}_{k1}^*\hat{X}_{h1}^*}$ represents the estimated joint probability of concurrent k-th and h-th attributes, i.e., $p^{\hat{X}_{k1}^*\hat{X}_{h1}^*} = p(\hat{X}_k^* = 1, \hat{X}_h^* = 1)$. Similarly, $p^{\hat{X}_{k0}^*\hat{X}_{h1}^*} = p(\hat{X}_k^* = 0, \hat{X}_h^* = 1)$. We define the corresponding penalty for the correlated attribute pair including positive and negative relations as follows

$$\ell(\hat{X}_k^*, \hat{X}_h^*) = \begin{cases} [p^{\hat{X}_{k1}^*}p^{\hat{X}_{h1}^*} - p^{\hat{X}_{k1}^*\hat{X}_{h1}^*}]_+ + [p^{\hat{X}_{k0}^*}p^{\hat{X}_{h1}^*} - p^{\hat{X}_{k1}^*\hat{X}_{h1}^*}]_+ \\ + [p^{\hat{X}_{k1}^*\hat{X}_{h0}^*} - p^{\hat{X}_{k1}^*\hat{X}_{h1}^*}]_+, \forall (k,h) \in U_P \\ [p^{\hat{X}_{k1}^*}p^{\hat{X}_{h1}^*} - p^{\hat{X}_{k1}^*\hat{X}_{h1}^*}]_+ + [p^{\hat{X}_{k1}^*}p^{\hat{X}_{h1}^*} - p^{\hat{X}_{k0}^*\hat{X}_{h1}^*}]_+ \\ + [p^{\hat{X}_{k1}^*\hat{X}_{h1}^*} - p^{\hat{X}_{k1}^*\hat{X}_{h0}^*}]_+, \forall (k,h) \in U_N \end{cases} \tag{5.9}$$

where $p^{\hat{X}_{k0}^*}$, $p^{\hat{X}_{k1}^*}$, $p^{\hat{X}_{h0}^*}$, $p^{\hat{X}_{h1}^*}$ and $p^{\hat{X}_{k1}^*\hat{X}_{h1}^*}$ are computed as: $p^{\hat{X}_{k0}^*} = \frac{\sum_n \phi(\hat{X}_{nk}^*=0)}{N}$, $p^{\hat{X}_{k1}^*} = \frac{\sum_n \phi(\hat{X}_{nk}^*=1)}{N}$, $p^{\hat{X}_{h0}^*} = \frac{\sum_n \phi(\hat{X}_{nh}^*=0)}{N}$, $p^{\hat{X}_{h1}^*} = \frac{\sum_n \phi(\hat{X}_{nh}^*=1)}{N}$, $p^{\hat{X}_{k1}^*\hat{X}_{h1}^*} = \frac{\sum_n \phi(\hat{X}_{nk}^*=1)\phi(\hat{X}_{nh}^*=1)}{N}$, where $\phi(x) = 1$ if x is true. $[\cdot] = max(\cdot, 0)$. Eq. 5.9 represents a loss of joint attributes dependencies. Thus, the total probabilistic loss of attributes correlations is

$$\ell_{pro}(\hat{\mathbf{X}}^*) = \sum_{(k,h)\in U} \ell(\hat{X}_k^*, \hat{X}_h^*). \tag{5.10}$$

Adversarial Loss

Inspired by the framework of GANs [Goodfellow et al., 2014], we explore attributes dependencies from ground truth attributes by adversarial learning. We introduce a discriminator $D(\mathbf{X_i}^*; \theta_D)$ to distinguish the predicted attributes from real attributes. For the generator network $f(\mathbf{X_i})$, we minimize the error between the prediction and ground truth via regression loss, ranking constraint, similarity constraint and prior probabilities loss. Furthermore, we want the generative network to make attributes predictions which "fool" the discriminator, leading to a new adversarial learning objective. Under generator network and the discriminator compete with each other, the generator network will be expected to output predictions, which minimize the error. Specifically, the learning objective of the framework can be written

as follows:

$$\min_{\theta_f} \max_{\theta_D} \sum_{i=1}^{N} \sum_{k=1}^{d^*} [\log D(X_{ik}^*) + \log(1 - D(\hat{X}_{ik}^*))]. \tag{5.11}$$

Similar to the optimization procedure of GAN, the learning objective in Eq.5.11 cannot be optimized directly. The discriminator and the generator network are optimized alternately by fixing their opponents. The learning objectives of these two components are reformulated into loss functions. The loss function of the discriminator is shown in Eq.5.14.

$$\ell_{adv} = \sum_{i=1}^{N} \sum_{k=1}^{d^*} [-\log D(X_{ik}^*) - \log(1 - D(\hat{X}_{ik}^*))]. \tag{5.12}$$

According to the suggestion in [Goodfellow, 2016], it is better for generative network to minimize $-\log D(\hat{X}_{ik}^*)$ instead of $\log(1 - \mathrm{D}(\hat{\mathrm{X}}_{\mathrm{ik}}^*))$ in order to avoid the flat gradient.

We combine the aforementioned guidance including Eq.5.1, 5.2, 5.10 and 5.14, and define the overall loss function of our proposed framework as follows:

$$\mathcal{L}(\theta_f) = \ell_{reg} + \alpha_1 \ell_{rank} + \alpha_2 \ell_{sim} + \alpha_3 \ell_{pro} + \alpha_4 \ell_{adv}, \tag{5.13}$$

where α_1, α_2, α_3 and α_4 are weights in order to balance the tradeoff among losses. Therefore, the optimization procedure of the proposed method is similar to the original GAN framework [Goodfellow et al., 2014]. We outline the learning procedure of the proposed method in Algorithm 4.

5.3.2 Partially Attribute-annotated Photos: Semi-supervised Adversarial Learning for Attribute-aware Photo Aesthetic Assessment

Aesthetic attributes are crucial for aesthetics because they explicitly present some photo quality cues that a human expert might use to evaluate a photo's aesthetic quality. However, annotating aesthetic attributes is a time-consuming, costly, and error-prone task, which leads to the issue that photos available are partially annotated with attributes. To alleviate this issue, we propose a novel semi-supervised adversarial learning method for photo aesthetic assessment from partially attribute-annotated photos, which can greatly reduce the reliance on manual attribute annotation. Specifically, the proposed method consists of a score-attributes generator R, a photo generator G, and a discriminator D. The score-attributes generator learns the aesthetic score and attributes simultaneously to capture their dependencies and

Algorithm 1 Model learning with PI guidance.

Require:
training sample $S = \{\mathbf{X}, \mathbf{X}^*, \mathbf{Y}\}$,
learning rate η_1 and η_2
the number of steps of updating adversarial network K_1, the number of steps of updating generative networks $K2$, bitch size m,
Ensure: the generative network f
Randomly initialize parameters of model parameters θ_D and θ_f;
for number of training iterations **do**
for $K1$ steps **do**
Sample a mini-batch of m training data $\{\mathbf{X_i}\}_{i=1}^m$ from $\{1, ..., n\}$
Perform forward propagation with $\{\mathbf{X_i}\}_{i=1}^m$ and fetch $\{\hat{\mathbf{X}}_\mathbf{i}^*\}_{i=1}^m$
Sample a mini-batch of m privileged information labels $\{\mathbf{X_i}^*\}_{i=1}^m$ from $\{1, ..., n\}$
Update the adversarial network D by gradient descent $\theta_D := \theta_D - \eta_1 \frac{\partial \ell_{adv}(\theta_D)}{\partial \theta_D}$
for $K2$ steps **do**
Sample a mini-batch of m training data $\{\mathbf{X_i}, \mathbf{X}_\mathbf{i}^*, \mathbf{Y_i}\}_{i=1}^m$ from $\{1, ..., n\}$
Perform forward propagation with $\mathbf{X_i}_{i=1}^m$ and fetch $\{\hat{\mathbf{X}}_\mathbf{i}^*\}_{i=1}^m$ and $\{\hat{\mathbf{Y}}_\mathbf{i}\}_{i=1}^m$
update the generative network f by gradient descent $\theta_f := \theta_f - \eta_2 \frac{\partial \mathcal{L}(\theta_f)}{\partial \theta_f}$

construct better feature representations. The photo generator reconstructs the photo by feeding aesthetic attributes, score, and informative feature representation. For attribute-annotated training data, we reconstruct the photo with aesthetic attributes, score from ground truth, and informative feature representation x_R from the penultimate layer features of R, then generate fake aesthetic attributes and score using the output of the photo generator as the input of the score-attributes generator. For no attribute-annotated training data, we reconstruct the photo with the output of the score-attributes generator and x_R as the input of the photo generator. A discriminator is used to force the convergence of the features-attributes-score tuples generated from the score-attributes generator, the photo generator, and the ground-truth distribution in labeled data for training data.

The framework of the proposed method is summarized in Figure 5.3. Specifically, through the score-attributes generator R, we get the predicted aesthetic attributes and aesthetic score $\{\hat{a}, \hat{y}\}$. Similarly, we sample data $\{a, y\}$ from B along with random noise $z \sim p_z(z)$ as the input of photo generator G, then we get the generated feature vector $\hat{x}^{'}$. The predicted aesthetic attributes and score $\{\hat{a}, \hat{y}\}$ along with the penultimate layer features x_R of R are input into the photo generator G, and the output $\hat{x}$ is the reconstruction of x. Similarly, the generated feature $\hat{x}^{'}$ is input into score-attributes generator R, and the output of $\{\hat{a}^{'}, \hat{y}^{'}\}$ is the reconstruction of a and y. Decoding network is used to strengthen the penultimate layer features x_R holding

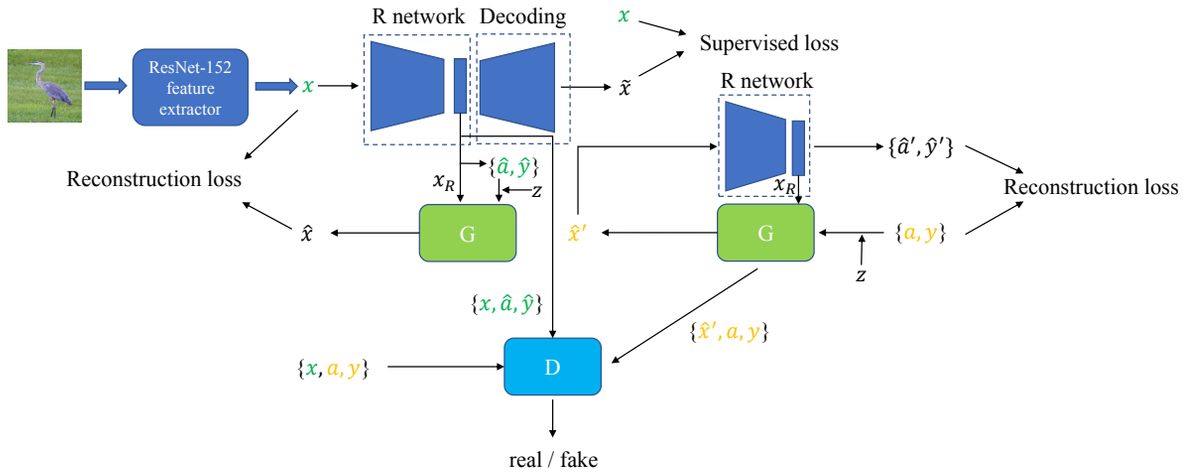


Figure 5.3: The framework of the proposed semi-supervised adversarial learning.

the information of photo features and promote G to generate better feature representations. Thus, we consider three kinds of losses in the following subsections. Then, we give the learning objective functions and optimization.

5.3.2.1 Adversarial Loss

As shown in Figure 5.3, the discriminator D tries to distinguish the tuples $\{x, \hat{a}, \hat{y}\}$ and $\{\hat{x}^{'}, a, y\}$ from the ground truth distribution of $\{x, a, y\}$ embedded in the training data. $\{x, \hat{a}, \hat{y}\}$ and $\{\hat{x}^{'}, a, y\}$ are generated by R and G respectively to "fool" D, $\{x, a, y\}$ are sampled from set T. Under the competing process among score-attribute generator, photo generator and discriminator, the R network will be expected to output predictions, which minimize the error. Specifically, the adversarial loss of discriminator is shown in Eq. 5.14.

$$\begin{aligned} \ell_{adv}^{D} = &- \mathbb{E}_{(x,a,y)\sim T} \log D(x, a, y) - \alpha \mathbb{E}_{(x,y)\sim V} \log(1 - D(x, \hat{a}, \hat{y})) \\ &- (1 - \alpha) \mathbb{E}_{(a,y)\sim B, z\sim p_z(z)} \log(1 - D(\hat{x}^{'}, a, y)). \end{aligned} \tag{5.14}$$

where $\alpha \in [0, 1]$ is a trade-off between the distribution of the pseudo-tuples generated from R and G network.

5.3.2.2 Supervised Loss

Supervised loss contributes to the capacity of the network to predict the commonly occurring range. As shown in Figure 5.3, we use the supervised loss in the generated attributes and score $\{\hat{a}, \hat{y}\}$ in R network. Supervised losses are also used in the generated features $\tilde{x}$ and $\hat{x}^{'}$ in decoding network and G network respectively. We

define them as:

$$\ell_{sl}^{R} = \mathbb{E}_{(x,a,y)\sim T} L(\{a,y\},\{\hat{a},\hat{y}\}), \tag{5.15}$$

$$\ell_{sl}^{Decoding} = \mathbb{E}_{(x,y)\sim V} L(x,\tilde{x}), \tag{5.16}$$

$$\ell_{sl}^{G} = \mathbb{E}_{(a,y)\sim B} L(x,\hat{x}^{'}), \tag{5.17}$$

where the loss function can be either squared error or binary cross entropy relying on the target variable is continuous or binary. The formulations of squared error and binary cross entropy are $L(y,\hat{y}) = (y-\hat{y})^2$ and $L(y,\hat{y}) = -y\log\hat{y} - (1-y)\log(1-\hat{y})$, respectively.

5.3.2.3 Reconstruction Loss

We introduce the reconstruction loss and hope that the G network and R network can reconstruct photo information, aesthetic attributes and aesthetic score. The reconstruction losses for network G and R are shown in the following:

$$\ell_{rl}^{G} = \mathbb{E}_{(x,y)\sim V, z\sim p_z(z)} L(x,\hat{x}), \tag{5.18}$$

$$\ell_{rl}^{R} = \mathbb{E}_{(a,y)\sim B} L(\{a,y\},\{\hat{a}^{'},\hat{y}^{'}\}), \tag{5.19}$$

where the loss function L is squared error or binary cross entropy depending on the continuous or binary type of target variable.

5.3.2.4 The Learning Objectives of Each Network

We give the objective functions of D, R and G networks, respectively.

Gradient descent on D network

$$\begin{aligned}\min_{\theta_D}[&-\mathbb{E}_{(x,a,y)\sim T}\log D(x,a,y) - \alpha\mathbb{E}_{(x,y)\sim V}\log(1-D(x,\hat{a},\hat{y})) \\ &-(1-\alpha)\mathbb{E}_{(a,y)\sim B, z\sim p_z(z)}\log(1-D(\hat{x}^{'},a,y))].\end{aligned} \tag{5.20}$$

Gradient descent on R network

$$\min_{\theta_R}[-\mathbb{E}_{(x,y)\sim V}\log(D(x,\hat{a},\hat{y})) + \lambda_{sr}\mathbb{E}_{(x,a,y)\sim T} L(\{a,y\},\{\hat{a},\hat{y}\}) + \lambda_{rr}\mathbb{E}_{(a,y)\sim B} L(\{a,y\},\{\hat{a}^{'},\hat{y}^{'}\})] \tag{5.21}$$

where λ_{sr} and λ_{rr} are weights of supervised loss and reconstruction loss respectively, in order to balance the tradeoff among losses.

Algorithm 2 The training of semi-supervised adversarial learning of attributes in aesthetic assessment

Require: training sample T, V and B, hyper parameters α, λ_{sr}, λ_{rr}, λ_{sg}, and λ_{rg}, bitch size c; the number of steps of updating discriminator network K_1, the number of steps of updating generator networks K_2, the number of training iterations K.

Ensure: score-attributes generator R and photo generator G.

Randomly initialize model parameters θ_D, θ_R and θ_G respectively;

for K iterations **do**

for K_1 steps **do**

Sample a mini-batch of c training data $\{x_i, a_i, y_i\}_{i=1}^{c}$ from T, sample a mini-batch of c training data $\{x_i, y_i\}_{i=1}^{c}$ from V, sample a mini-batch of c training data $\{a_i, y_i\}_{i=1}^{c}$ from B.

Update discriminator D by gradient descent:

$$\nabla_{\theta_D}[-\frac{1}{c}\sum_{i=1}^{c}((\log D(x,a,y)+\alpha\log(1-D(x,\hat{a},\hat{y}))+(1-\alpha)\log(1-D(\hat{x}^{'},a,y)))]$$

for K_2 steps **do**

Sample a mini-batch of c training data $\{x_i, y_i\}_{i=1}^{c}$ from V, c_1 ($c_1 \leq c$) samples of $\{x_i, y_i\}_{i=1}^{c}$ originally annotated with aesthetic attributes are used to construct $\{x_i, a_i, y_i\}_{i=1}^{c_1}$.

Update score-attribute generator R by gradient descent:

$$\nabla_{\theta_R}[-\frac{1}{c}\sum_{i=1}^{c}\log(D(x,\hat{a},\hat{y}))+\frac{\lambda_{sr}}{c_1}\sum_{i=1}^{c_1}L(\{a,y\},\{\hat{a},\hat{y}\})+\frac{\lambda_{rr}}{c_1}\sum_{i=1}^{c_1}L(\{a,y\},\{\hat{a}^{'},\hat{y}^{'}\})]$$

Sample a mini-batch of c training data $\{a_i, y_i\}_{i=1}^{c}$ from B, sample a mini-batch of c training data $\{x_i, y_i\}_{i=1}^{c}$ from V.

Update photo generator G by gradient descent:

$$\nabla_{\theta_G}[-\frac{1}{c}\sum_{i=1}^{c}\log(D(\hat{x}^{'},a,y))+\frac{\lambda_{sg}}{c}\sum_{i=1}^{c}L(x,\hat{x}^{'})+\frac{\lambda_{rg}}{c}\sum_{i=1}^{c}L(x,\hat{x})]$$

Gradient descent on G network

$$\min_{\theta_G}[-\mathbb{E}_{(a,y)\sim B,z\sim p_z(z)}\log(D(\hat{x}^{'},a,y))+\lambda_{sg}\mathbb{E}_{(a,y)\sim B}L(x,\hat{x}^{'})+\lambda_{rg}\mathbb{E}_{(x,y)\sim V,z\sim p_z(z)}L(x,\hat{x})] \tag{5.22}$$

where λ_{sg} and λ_{rg} are weights of supervised loss and reconstruction loss respectively.

The optimization of the proposed objective function is similar to the original GAN framework [Goodfellow et al., 2014]. We outline the proposed algorithm in Algorithm 4.

5.4 Experiments

5.4.1 Experimental Conditions

5.4.1.1 Datasets

Table 5.1: Attribute annotations on the AADB and AVA database.

Attribute (AADB)	Number	Attribute (AVA)	Number
Balancing elements	10,000	Complementary	949
Color Harmony	10,000	Duotones	1,301
Content	10,000	HDR	396
Depth of Field	10,000	Photo Grain	840
Light	10,000	Light on White	1,199
Motion Blur	10,000	Long Exposure	845
Object	10,000	Macro	1,698
Repetition	10,000	Motion Blur	609
Rule of Thirds	10,000	Negative Blur	959
Symmetry	10,000	Rule of Thirds	1,031
Vivid Color	10,000	Shallow DoF	710
-	-	Silhouettes	1,389
-	-	Soft Focus	1,479
-	-	Vanishing Point	674

In our experiments, we use two publicly available photo aesthetics databases: Aesthetics and Attributes database (AADB) [Kong et al., 2016] and Aesthetics Visual Analysis database (AVA) [Murray et al., 2012]. AADB database contains 10,000 photographic images from the Flickr website [Kong et al., 2016]. Aesthetic quality scores are continuous values in the interval of [0, 1]. The AADB database provides 11 attributes that are continuous values including balancing element, content, color harmony, depth of field, lighting, motion blur, object emphasis, rule of third, vivid color, repetition and symmetry. The official split of the AADB database is 8,500 images for training, 500 images for validation and the remaining 1,000 images for testing. The AVA database contains aesthetic quality scores in the interval [1, 10] and 14 style attributes in a small portion of photos (14,000/250,000), i.e., complementary, duotones, HDR, photo grain, light on white, long exposure, macro, motion blur, negative photo, rule of third shallow DOF, silhouettes, soft focus and vanishing point. These attributes are binary with discrete values of 0 or 1. Moreover, the AVA database is split into training set (230,000 samples) and testing set (20,000 samples). The official split of the AVA database does not include a validation set, and thus we have to construct the validation set by ourselves. We randomly sample 20,000 photos from the training set as the validation set, and the remaining 210,000 photos for training.

The detailed attributes annotations on the AADB and AVA databases are listed in Table 5.1.

5.4.1.2 Implementation Details

Implementation Details of Multi-task Learning Network

In our experiments, photos are resized to 224×224 on both AADB and AVA databases. We use ResNet-152 [He et al., 2016b] as the architecture of the proposed network to implement our method. The pre-trained weights of ResNet-152 on Imagenet are used for initialization. Then 2048D feature representations are extracted from this pre-trained architecture. Afterwards, we build the three hidden fully connected layers with ReLU activations. The sizes of these three layers are 512, 128 and (1+d^*) respectively where d^* depends on actual attributes dimensions on databases. The attributes on AADB are 11 dimensions with continuous values normalized to interval of [0, 1], while the attributes on AVA are 14 dimension with binary of 0 or 1. The aesthetic scores on two databases are normalized to the interval of [0,1]. On discriminator network, we build a three-layer neural network with a size of (1+d^*)×8×1. We use Adam [JLB, 2015] optimizer with a mini-batch size of 64. The number of algorithm runs is 200. The learning rates of the generator network and discriminator are 1e-4 and 5e-5 respectively. The weight decay is set to be 1e-5. $\alpha_1 = 0.001$, $\alpha_2 = 0.02$, $\alpha_3 = 0.005$ and $\alpha_4 = 0.01$ and and $C_k(1 \leq k \leq d^*)$ are [0.340, 0.569, 0.697, 0.416, 0.574, 0.204, 0.542, 0.359, 0.518, 0.620, 0.470] among these 11 kinds of attributes on AADB, and [0.682, 0.540, 0.510, 0.620, 0.473, 0.425, 0.664, 0.432, 0.466, 0.374, 0.421, 0.383, 0.460, 0.370] among these 14 kinds of attributes on AVA.

Implementation Details of Semi-supervised Adversarial Learning Network

We implement our method using the PyTorch framework. In our experiments, photos are resized to 224×224 on both databases. We normalize the aesthetic score and continuous attributes to the interval [0, 1]. ResNet-152 [He et al., 2016b] is used as the architecture to extract the feature representations. The pre-trained weights of ResNet-152 on Imagenet are used for initialization. Then 2048D feature representations are extracted from this pre-trained architecture. Score-attributes generator R, photo generator G, discriminator D and decoding network are parameterized by four-layer feed forward networks. Adam [JLB, 2015] optimizer with a mini-batch size of 64 is used to optimize R, G, D and decoding network in our experiments. Other hyper parameters, such as learning rate, training step K_1, K_2, and weight coefficients λ_{sr}, λ_{rr}, λ_{sg}, λ_{rg} are determined by a validation set.

5.4.1.3 Evaluation Metric

The Spearman Rank-order Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC) [Hosu et al., 2019], accuracy [Kong et al., 2016] and Mean-Square Error (MSE) are used as metric evaluation. In terms of SRCC, PLCC and accuracy, the larger the better and for MSE, the lower the better.

5.4.1.4 Baselines

Firstly, we compare our method to attribute-aware methods. We choose the following eight popular methods for comparison. For a fair comparison, we re-implement RA-DCNN method by changing AlexNet [Kong et al., 2016] with ResNet framework.

- Ranking and attribute deep convolutional rating network(RA-DCNN) includes additional activation layers in their proposed ranking network to encode informative attributes. Their methods regard prediction task as a source of side-information to regularize the weights of the network [Kong et al., 2016].
- Attribute-aware deep convolution neural network (DCNN) [Malu et al., 2017] jointly learns aesthetic attributes and aesthetic score simultaneously to learn the inherent representation of these attribute, and qualitatively analyze the diverse and complex association with these different attributes for aesthetic assessment.
- Adversarial deep convolutional rating network (ATTGAN) [Pan et al., 2019a] proposes using deep rating network as a generator outputting attributes and score simultaneously. ATTGAN uses a discriminator to distinguish the generative attributes, score and ground truth by regularizing the distributions of high-level attributes and score.
- multi-task deep learning (MTDL) [Li et al., 2019] proposes to use the personality features that are learned to modulate the aesthetic attributes. These attributes are used to predict the optimal generic photo aesthetics scores.
- Privileged information deep convolutional network (PI-DCNN) [Shu et al., 2020] uses attribute loss function, which explores the domain knowledge of the photo-based and photography-based attributes to improve the performance of aesthetic assessment.
- Support vector regression with variational privileged information(v-SVR+) [Shu et al., 2021] integrates aesthetic attributes as privileged information to predict the slack variable of support vector regression further regularizing the learning process and improving the learning performance of aesthetic prediction.

- Deep chatterjee's machine (DCM) [Wang et al., 2017] first learns aesthetic attributes via a parallel supervised pathway. Then they associate and transform those attributes into the overall aesthetic quality assessment.
- A multi-task framework (MTRLCNN) [Kao et al., 2017] is proposed to capture the correlations between semantic recognition and aesthetic quality prediction via incorporating the inter-task relationship learning.

Secondly, we compare our method to attribute-unaware methods. We choose the following seven popular methods for comparison.

- Squared earth mover's distance (Square-EMD) [Hou et al., 2016] is used to define loss functions for convolutional neural network training. They transform aesthetic database into a classification database, and leverage these relations between different classes via the squared earth mover's distance.
- A unified probabilistic formulation (AUPF) [Zeng et al., 2019] uses an effective loss function to capture the inherent relation among binary classification, average score regression and score distribution prediction.
- Composition-aware network(CAN) [Liu et al., 2020] leverages the image composition information as the mutual dependency among its local regions to boost the performance of aesthetics.
- Convolutional ordinal regression forest (CORF) [Zhu et al., 2021] obtains precise and stable global ordinal relationships via integrating ordinal regression and differentiable decision trees with a CNN for photo aesthetic assessment.
- Neural image assessment (NIMA) model [Talebi and Milanfar, 2018] uses convolutional neural network to predict the distribution of human opinion scores which capture the high correlations between photo score and human perception.
- Gated peripheral-foveal convolutional neural networks (GPF-CNN) [Zhang et al., 2019] encodes the holistic information to provide the attended regions, then extracts fine-grained features on these key regions.
- Multi-level spatially pooled features (MLSP) [Hosu et al., 2019] are extracted from convolutional blocks to efficiently support full resolution images on variable input sizes.

5.4.1.5 Experiment Design on Full Attribute-annotated Photo Aesthetic Assessment

We evaluate our method with different combinations of loss functions on the AADB and AVA databases. The first one is the method of BesNet. BesNet refers to Basic network in Sec 5.3.1.1, which uses a standard ResNet-152 with regression loss ℓ_{reg} for basic network. The second one is BesNet with ℓ_{rank} in label space. The third one is the method of basic network with ℓ_{sim}, ℓ_{pro}, ℓ_{adv} and combination of them in privileged information space. The fourth one is the proposed method which combines ℓ_{reg}, ℓ_{rank}, ℓ_{sim}, ℓ_{pro} and ℓ_{adv}.

5.4.1.6 Experiment Design on Partially Attribute-annotated Photo Aesthetic Assessment

The number of attributes-annotated photos on the AADB and AVA database is 10,000 and 14,000, respectively. We randomly exclude attributes-annotated photos with four ratios, i.e., 10%, 20%, 30% and 40%. The proposed method employs aesthetic attributes as privileged information [Vapnik and Vashist, 2009] to improve the learning performance of aesthetic assessment tasks. We compare the proposed method to a method that does not output aesthetic attributes in R network, referred to as SGAN. Besides, we compare the proposed method with another method that removes G network in our method, referred to as SAGAN-.

5.4.2 Results of Within-Database Experiments

5.4.2.1 Fully Attribute-annotated Results

Table 5.2: Experimental results on the AADB and AVA databases.

Model	AADB	Model	AVA		
	SRCC		SRCC	Accuracy	PLCC
RA-DCNN	0.6782	RA-DCNN	0.5581	77.30%	0.563
Square-EMD	0.6889	NIMI	0.6120	-	0.636
DCNN	0.6890	GPF-CNN	0.6900	81.81%	0.704
MTDL	0.6800	MLSP	0.756	81.72%	0.757
ATTGAN	0.7041	ATTGAN	0.6313	-	0.652
PI-DCNN	0.7051	V-SVR+	0.6880	80.43%	0.691
Ours	**0.7670**	**Ours**	**0.7580**	**83.48%**	**0.768**

Table 5.2 shows the experimental results of all comparison methods on the AADB and AVA database. Compared with these methods, our method uses aesthetic attributes guidance in privileged information space and label space. Specifically, our

Table 5.3: Within database experimental results of semi-supervised adversarial learning for aesthetic assessment with four missing rates of attribute-annotated photos on the AADB and AVA databases.

Database	Methods	10%			20%			30%			40%		
		SRCC	Accuracy	PLCC	SRCC	Accuracy	PLCC	SRCC	Accuracy	PLCC	SRCC	Accuracy	PLCC
AADB	v-SVR+	0.684	79.62%	0.676	0.662	77.55%	0.667	0.638	75.43%	0.640	0.620	74.09%	0.629
	RA-DCNN	0.665	77.54%	0.658	0.643	75.88%	0.644	0.621	74.64%	0.627	0.603	72.18%	0.597
	PI-DCNN	0.689	79.96%	0.686	0.674	78.02%	0.672	0.633	76.25%	0.628	0.602	72.68%	0.616
	ATTGAN	0.687	78.42%	0.692	0.674	77.65%	0.678	0.646	74.83%	0.652	0.605	72.11%	0.611
	SGAN	0.686	78.88%	0.683	0.664	76.52%	0.674	0.643	74.40%	0.655	0.635	73.69%	0.639
	SAGAN-	0.692	79.52%	0.702	0.677	78.13%	0.682	0.653	76.04%	0.660	0.648	75.52%	0.647
	SAGAN	**0.747**	**81.47%**	**0.752**	**0.722**	**80.56%**	**0.737**	**0.703**	**79.94%**	**0.714**	**0.695**	**79.18%**	**0.705**
AVA	v-SVR+	0.677	78.52%	0.665	0.662	76.43%	0.668	0.642	72.28%	0.651	0.582	68.84%	0.574
	RA-DCNN	0.547	74.62%	0.558	0.556	72.15%	0.534	0.538	70.11%	0.526	0.529	68.42%	0.517
	PI-DCNN	0.663	77.62%	0.676	0.638	73.48%	0.652	0.619	71.47%	0.628	0.592	69.93%	0.604
	ATTGAN	0.628	76.44%	0.642	0.608	74.18%	0.617	0.596	72.26%	0.604	0.587	70.04%	0.593
	SGAN	0.628	75.42%	0.635	0.614	73.71%	0.624	0.598	72.63%	0.605	0.584	69.18%	0.592
	SAGAN-	0.633	76.42%	0.638	0.614	74.68%	0.621	0.601	72.13%	0.598	0.587	70.07%	0.592
	SAGAN	**0.750**	**82.16%**	**0.758**	**0.734**	**81.03%**	**0.742**	**0.717**	**80.66%**	**0.722**	**0.695**	**79.42%**	**0.704**

proposed method uses prior probabilities loss to capture the dependencies among multiple attributes and introduces adversarial learning to force the distribution of the prediction and ground truth to be closer. Therefore, our method achieves the best Spearman rank-order correlation coefficient, Pearson linear correlation coefficient, and accuracy.

5.4.2.2 Partially Attribute-annotated Results

Analyses of Four Missing Rates

Table 5.3 demonstrates the within-database experimental results of semi-supervised attribute learning on the AADB and AVA databases. Among these methods, (i) v-SVR+ is the work using attributes on support vector regression; (ii) ATTGAN, SGAN and SAGAN- are the works of GAN variations; (iii) RA-DCNN and PI-DCNN are the works applying attributes on deep convolutional neural network.

From Table 5.3, we make the following observations: First, compared to v-SVR+, RA-DCNN, PI-DCNN and ATTGAN, the proposed SAGAN has the best performance. For example, with the 40% missing rate, the performance of SAGAN are 7.5%/5.1%/7.6%, 9.2%/7.0%/10.8%, 9.3%/6.5%/8.9% and 9.0%/7.1%/13.9% higher SRCC/Accuracy/PLCC than v-SVR+, RA-DCNN, PI-DCNN and ATTGAN on the AADB database. Similarly significant improvement can be found on the AVA database. This demonstrates the superiority of the proposed method due to more attributes generated in SAGAN. Second, SAGAN performs better in all scenarios compared to SGAN that does not output aesthetic attributes on R network. The reason is that SAGAN learns the aesthetic attributes and score simultaneously to capture their dependencies and extract better feature representations. Third, when the missing rate varies from 10% to 40%, the gap performance between SAGAN and the other methods is becoming large. This is because the proposed method can generate more



Figure 5.4: Examples of predicted aesthetic scores on the AADB database: RA-DCNN/ATTGAN/Ours/Ground truth.

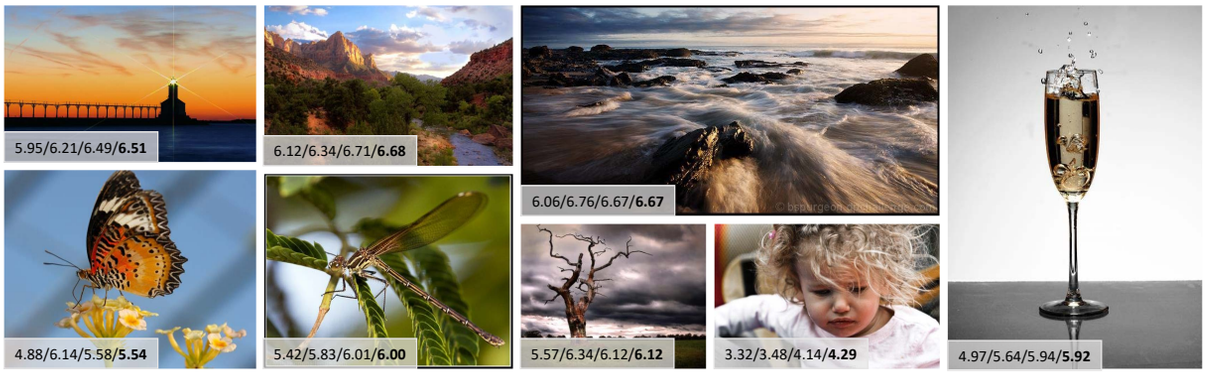


Figure 5.5: Examples of predicted aesthetic scores on the AVA database: RA-DCNN/ATTGAN/Ours/Ground truth.

representative aesthetic attributes after missing, which is more beneficial for photo aesthetic assessment. Fourth, compared to SAGAN- that removes photo generator G from our method, SAGAN has a better performance in all scenarios. This is attributed to the fact that photo generator G leveraged to reconstruct the generated feature vector is important improvement for our proposed method.

Comparison of the Predicted Photo Examples

Figure 5.4 and Figure 5.5 show the predicted aesthetic scores of our methods compared to two attribute-aware methods (RA-DCNN and ATTGAN) and ground

truth. We select some photos from the testing set of the AADB and AVA databases, respectively. Generally, we observe that our method has a lower prediction error compared to RA-DCNN and ATTGAN. The slight gap between our predicted scores and ground truth indicates that our trained models can accurately predict aesthetic scores.

Comparison to State-of-the-art Attribute-Aware Methods

Table 5.4: Comparison to state-of-the-art attribute-aware methods on the AADB and AVA database.

Model	AADB	Model	AVA		
	SRCC		SRCC	Accuracy	PLCC
RA-DCNN	0.678	RA-DCNN	0.558	77.03%	0.563
DCNN	0.689	DCM	-	78.08%	-
MTDL	0.680	MTRLCNN	-	79.08%	-
ATTGAN	0.704	ATTGAN	0.631	77.23%	0.655
PI-DCNN	0.705	PI-DCNN	0.658	76.2%	0.672
SAGAN	**0.761**	**SAGAN**	**0.774**	**83.72%**	**0.788**

To further evaluate the performance of the proposed method, we compare the proposed semi-supervised attributes learning method to related attribute-aware methods shown in Table 5.4. From Table 5.4, we find that on the AADB and AVA database, the proposed SAGAN method has the best performance w.r.t. the highest SRCC, accuracy and PLCC than the other methods. Compare to these methods using the same number of attributes that both databases provide, the proposed SAGAN method benefits both the score-attributes generator and photo generator tasks via generating informative signals. Compared to state-of-the-art attribute-aware methods that heavily rely on manual attributes annotation, the proposed method uses only partially attribute-annotated photos and achieves the best performance of aesthetic prediction.

Comparison to State-of-the-art Attribute-Unaware Methods

Table 5.5: Comparison to state-of-the-art attribute-unaware methods on the AADB and AVA databases.

Model	AADB	Model	AVA		
	SRCC		SRCC	Accuracy	PLCC
Square-EMD	0.689	NIMI	0.612	-	0.636
AUPF	0.726	GPF-CNN	0.690	81.81%	0.704
CAN	0.710	MLSP	0.756	81.72%	0.757
CORF	0.677	CORF	0.671	-	0.665
Ours	**0.761**	**Ours**	**0.774**	**83.72%**	**0.788**

To further demonstrate the importance of high-level attributes in aesthetic assessment, we compare our method to state-of-the-art attribute-unaware methods shown in Table 5.5. As we can see from Table 5.5, the proposed method using high-level attribute in the aesthetic rating network achieves the best performance w.r.t. the highest SRCC, accuracy and PLCC than the other methods, which demonstrates that exploiting high-level aesthetic attribute can significantly benefit aesthetic assessment task and our proposed method successfully leverages the correlation between aesthetic attributes and aesthetic score.

5.4.3 Results of Cross-Database Experiments

5.4.3.1 Fully Attribute-annotated Results

Table 5.6: SRCC (Spearman Rank-order Correlation Coefficient) for training and testing on Cross-Database.

Training set	AADB		AVA	
Testing set	AADB	AVA	AADB	AVA
RA-DCNN	0.6782	0.1566	0.3191	0.5581
NIMI	0.7081	0.2950	0.4732	0.6904
ATTGAN	0.7041	0.3104	0.4582	0.6313
AUPF	0.7264	0.3225	0.5176	0.7141
Ours	**0.7670**	**0.3462**	**0.6052**	**0.7580**

The results of cross-database experiments are shown in Table 5.7. When the methods are trained on the AADB database and tested on the AVA database, our method achieves the best performance. Similarly, when the methods are trained on the AVA and tested on the AADB, our method substantially improves the cross-database evaluation performance and is superior to the other methods. Notably, the improvement of our method from AVA to AADB is more obvious than our method from AADB to AVA. The reason is that aesthetic attributes on the AVA with 14 attributes is trained better than on the AADB with 11 attributes.

5.4.3.2 Partially Attribute-annotated Results

Analyses of Four Missing Rates

The cross-database experimental results of semi-supervised adversarial learning for attribute-aware photo aesthetic assessment task on the AADB and AVA databases are shown in Table 5.7. From Table 5.7, we find: First, when the methods are trained on the AADB database and tested on the AVA database, the proposed SAGAN

Table 5.7: Cross-database evaluation of semi-supervised adversarial learning for aesthetic assessment with four missing rates of attribute-annotated photos.

Database	Methods	10%			20%			30%			40%		
		SRCC	Accuracy	PLCC	SRCC	Accuracy	PLCC	SRCC	Accuracy	PLCC	SRCC	Accuracy	PLCC
From AADB to AVA	v-SVR+	0.254	59.88%	0.261	0.232	58.43%	0.248	0.206	55.42%	0.217	0.192	53.39%	0.202
	RA-DCNN	0.182	57.09%	0.197	0.173	55.62%	0.179	0.154	53.37%	0.162	0.143	52.03%	0.150
	PI-DCNN	0.274	60.57%	0.271	0.253	59.98%	0.248	0.221	57.72%	0.236	0.205	56.43%	0.217
	ATTGAN	0.263	60.02%	0.270	0.241	59.34%	0.254	0.226	58.52%	0.238	0.217	57.79%	0.224
	SGAN	0.243	58.77%	0.250	0.232	56.40%	0.244	0.225	55.38%	0.228	0.210	55.02%	0.214
	SAGAN-	0.259	59.63%	0.263	0.234	58.02%	0.240	0.225	57.36%	0.232	0.209	56.48%	0.205
	SAGAN	**0.327**	**68.82%**	**0.334**	**0.306**	**66.47%**	**0.312**	**0.288**	**62.32%**	**0.285**	**0.280**	**61.18%**	**0.278**
From AVA to AADB	v-SVR+	0.486	68.94%	0.498	0.463	66.43%	0.471	0.445	64.12%	0.450	0.417	62.39%	0.412
	RA-DCNN	0.332	67.54%	0.340	0.313	64.39%	0.311	0.298	60.12%	0.304	0.283	59.33%	0.293
	PI-DCNN	0.493	69.42%	0.502	0.477	67.99%	0.480	0.442	64.33%	0.447	0.394	63.50%	0.405
	ATTGAN	0.506	70.62%	0.511	0.488	68.54%	0.492	0.463	66.12%	0.268	0.432	63.12%	0.444
	SGAN	0.492	69.11%	0.498	0.477	68.12%	0.482	0.453	66.53%	0.451	0.442	64.88%	0.437
	SAGAN-	0.502	70.58%	0.506	0.487	69.02%	0.492	0.465	67.66%	0.472	0.450	65.32%	0.448
	SAGAN	**0.583**	**74.43%**	**0.601**	**0.565**	**72.12%**	**0.577**	**0.554**	**70.05%**	**0.561**	**0.547**	**69.32%**	**0.552**

achieves the best performance in the different missing rates of attributes. For example, with the 40% missing rate, the performance of SAGAN are 8.8%/7.8%/7.6%, 13.7%/9.15%/12.8%, 7.5%/4.8%/6.1% and 6.8%/3.4%/5.4% higher SRCC/Accuracy/PLCC than the method of v-SVR+, RA-DCNN, PI-DCNN and ATTGAN. Second, the SGAN perform better than RA-DCNN in all cases. Although SGAN is not assisted by attributes in R network, it takes advantage of the competition between photo synthesis and aesthetic score assessment, hence still achieves the better performance than RA-DCNN. Third, compared to the method of SAGAN- that removes photo generator G, the method of SAGAN perform better, which demonstrate photo generator plays an important roles in regularizing the score-attributes generator. A similar conclusion can be found on the AVA database.

Comparison to the State-of-the-art Methods

Table 5.8: Cross-database evaluation of SRCC reported on current studies.

Model	From AADB to AVA	From AVA to AADB
RA-DCNN	0.1566	0.3191
NIMA	0.2950	0.4732
AUPF	0.3225	0.5176
SAGAN	**0.3462**	**0.6052**

Table 5.8 shows the related cross-database experimental results on the AADB and AVA databases. Compared to current studies on cross database, the proposed method has a better performance when training on the AADB testing on the AVA and training on the AVA testing on the AADB. This demonstrates that the proposed method further explores the adaptation of different photo groups and photo collections. Note that cross-database performance is worse than within-database performance because the AADB database contains photos with different distributions of visual characteristics compared to the AVA database. Photos on the AVA database are professionally photographed or heavily edited; while many photos in the AADB

database are related to many daily life. Thus, the generalization of the proposed method should be further improved in the future work.

5.5 Discussion and Conclusion

In this chapter, we propose a deep multi-task network that utilizes high-level descriptive aesthetic attributes as privileged information to enhance the photo assessment. Pairwise photos are first trained respectively via a basic generator network. During the fine-tuning process of generative network, the ranking constraint is introduced in the label space to capture the ordinal relation between attributes and aesthetic score. In privileged information space, adversarial learning is used to enforce similarity between join distribution of the predicted attributes and ground truth attributes; prior probabilities loss is used to exploit the coexistent and exclusive relations among attributes; and similarity constraint are introduced to further regularize the learned aesthetic attributes. Our method achieves state-of-the-art performance on the task of photo aesthetic assessment on two benchmark databases, demonstrating that the proposed method captures the correlation between aesthetics and privileged information hidden in photos, and then can benefit the aesthetic assessment.

Furthermore, for partially attribute-annotated photos, we propose a semi-supervised learning method with the aim of reducing the reliance on manual attributes annotation for aesthetic assessment. An adversarial training framework is proposed to explore the joint distribution among photo features, aesthetic attributes and aesthetic score. Aesthetic attributes are used as privileged information to construct a better score-attribute generator and a photo generator. Supervised losses are used for the proposed networks predicting on commonly occurring range. Reconstruction losses are designed for regularizing the score-attributes generation and photo generation. Experimental results on the AADB and AVA databases demonstrate that the proposed method successfully leverages the inherent connections between aesthetic attributes and aesthetics through semi-supervised attributes learning and adversarial training process, and thus achieves the superior performance.

Chapter 6

Learning with Privileged Information for Music Emotion Recognition

6.1 Introduction

Digital music and online music services have become an essential part of modern daily life thanks to the availability and accessibility of the dramatically large amount of online digital music sources. Since music is often created to express human feelings, it is a wise way to retrieve and organize music by emotion learning. A typical approach to music emotion learning is to labelling the music emotion via tags in terms of semantic descriptions, and applies machine learning to train a classifier for automatically recognizing the emotion of a music piece. However, since manual labelling emotion tags for each music piece is time consuming, it is necessary to propose an intelligent model to recognize music emotion expressed automatically for music creation and music distribution. In this environment, naturally, automatic emotion recognition from music pieces has become an increasingly popular research topic in recent years.

Currently, research on music emotion recognition (MER) mainly focuses on feature engineering with existing traditional classification models. For example, various acoustic and psychoacoustic features including zero-crossing rate and mel-frequency cepstrum coeffcients (MFCC) etc. [Eyben et al., 2010], deep features extracted by deep learning models [Panda et al., 2018] or high-level semantic features (e.g., genre) [Friberg et al., 2014] are extracted and analyzed to identify the perceived emotion states of a music segment. Then, the music emotion recognition task can be modelled as a classification problem to classify music pieces into several discrete emotion categories, or it can be interpreted as a regression task that predicts continuous emotional di-

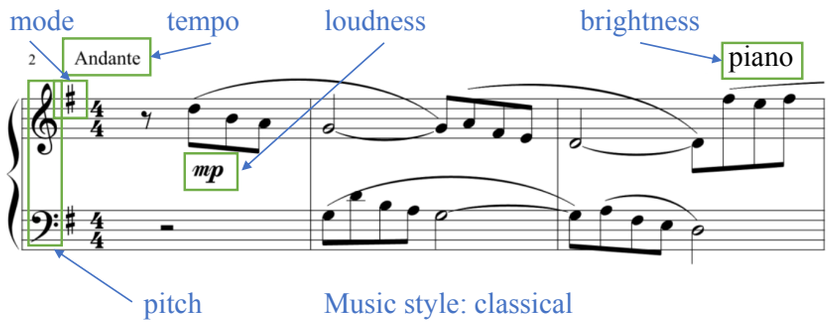


Figure 6.1: Example of music piece with prior information i.e. musical elements and music styles. This music piece is summarized from "The Butterfly Lovers", which is a classical music. From this music score, we can learn main music information, e.g. "Andante" represents musical tempo; "mp" represents musical loudness; Treble clef and bass clef indicate musical pitch; Sharp and flat refer to musical mode (See Figure 6.2); Different instruments result in different musical brightness [Saitis and Siedenburg, 2020]. These musical information collected in a laboratory environment are beneficial for MER, which can be used in training phase to improve the performance of model.

mensions, such as valence and arousal [Barrett, 1998]. In the early stages, researchers explored traditional classifiers such as support vector machines (SVMs) and support vector regression (SVR) [Liu et al., 2015; Chmulik et al., 2015; Weber et al., 2015], while in recent years, many studies have investigated deep learning based approaches, such as convolutional and recurrent neural networks [Malik et al., 2017].

The aforementioned research works on music emotion recognition focus on various audio features and classifiers. Compared with deep features which is opaque and difficult to interpret [Panda et al., 2020], perceptual features (e.g., rhythm clarity, tonality, timbre, etc.) and high-level semantic features (e.g., genre) [Friberg et al., 2014] improve the possibility to acquire novel knowledge regarding the relations between emotions and the extracted features. Such informative features can be exploited as the domain knowledge from musicology of emotion and music, i.e. the inherent psychological relationship between musical elements, music styles (genres) and human emotion. They are useful to construct better models during training by involving the structure or parameters of the classifier. For example, the main musical dimensions, i.e., rhythm, tonality, timbre, dynamics and pitch [Zbikowski and Zbikowski, 2002][Lartillot, 2011] which are hidden in songs indicate the possible emotional experience when audiences receive music messages. Such implicit cues can be leveraged for emotion recognition from songs; music styles (genres) including pop, classical, electronic etc. are also beneficial for music emotion recognition. Figure 7.1 shows the Chinese classical music piece "The Butterfly Lovers" [Jiang,

2004] where the information of musical elements and music styles are shown in music scores but hidden in the audio. Such important cues can be used as privileged information [Vapnik and Vashist, 2009] to provide some guidance and an explanation using classifiers. Typically, this privileged information is more informative for boosting the learning and recognition tasks [Sharmanska et al., 2013].

To the best of our knowledge, musical elements and music styles have not been exploited thoroughly in the current research. Current studies only regard musical elements as general training and testing feature sets [Hu and Yang, 2017][Chen et al., 2014], and the same for acoustic and psychoacoustic features (MFCC etc.). However, as high-level musicology features, musical elements represent domain knowledge in the music area, which can be used as privileged information to provide intuitive comments, comparisons, and explanations of classifiers [Lambert et al., 2018]. Similarly, current studies either ignore the domain knowledge of music styles or conduct experiments on a single category. For example, Hu *et al.* evaluated mood regression models built on fifteen acoustic features in five mood-related musical aspects only on pop songs [Hu and Yang, 2017]. Markov *et al* created two systems, one for music genre classification and another for music emotion estimation using both SVM and GP models. However, they regarded music genre classification, and music emotion estimation as two independent tasks but they fail to build the intrinsic relations between these two tasks [Markov and Matsui, 2014].

Therefore, in this chapter, we propose a novel element-enhanced and style-enhanced music emotion recognition, where the musical elements and music styles are used as privileged information. Our work aims to capture the intrinsic relations among target variable (i.e. emotion labels), available data (i.e. acoustic and psychoacoustic features), and privileged information (i.e. musical elements and music styles) to achieve a comparable level of emotion prediction. First, we use RBM to model the dependencies between musical elements and music styles. We divide the visible units of RBM into two parts, one representing the musical element features, and the other representing music style features. During training, both musical element features and music styles are used to learn the parameters of RBM models. After this, a new privileged information space integrating the musical elements and music styles can be obtained. Second, we introduce a privileged information-based support vector regression (SVR+) where we define a linear correcting (slack) function in the new privileged information space to estimate the slack variables in the standard SVR method using privileged information.

6.2 Summary of Domain Knowledge

Table 6.1: Analysis for emotional effects on musical elements and music styles. Note that High-A, Low-A, Positive-V and Negative-V represent high arousal, low arousal, positive valence and negative valence, respectively. $\surd$ denotes the existence of high dependency. $\uparrow$ denotes the associated emotions under improving the row status, $\downarrow$ denotes the associated emotions under descending the row status.

		H-A	L-A	P-V	N-V	Emotion
Dimension-related	Tempo	$\surd$	$\surd$			excitement $\uparrow$ peace $\downarrow$ anger $\uparrow$ sadness $\downarrow$
	Mode			$\surd$	$\surd$	happiness $\uparrow$ sadness $\downarrow$ sereneness $\uparrow$ anger $\downarrow$
	Brightness	$\surd$	$\surd$			excitement $\uparrow$ sadness $\downarrow$ astonishment $\uparrow$ softness $\downarrow$
	Loudness	$\surd$	$\surd$			excitement $\uparrow$ fear $\downarrow$ happiness $\uparrow$ tenderness $\downarrow$
	Pitch	$\surd$	$\surd$	$\surd$	$\surd$	anger $\uparrow$ pleasantness $\downarrow$ surprise $\uparrow$ boredom $\downarrow$
Style-related	Pop	$\surd$		$\surd$	$\surd$	dancing, happy, energetic, excitement
	Blues		$\surd$		$\surd$	solemnity, grace, sadness, dignity
	Classical		$\surd$	$\surd$		calmness, relaxing, peaceful, joy
	Country	$\surd$		$\surd$		happiness, excitement
	Electronic	$\surd$		$\surd$	$\surd$	excitement, powerful, disgust
	Folk	$\surd$			$\surd$	longing, dancing, romantic
	International	$\surd$	$\surd$	$\surd$	$\surd$	grace, elegance, calm
	Jazz	$\surd$		$\surd$		longing, amazement, spirituality, peacefulness
	Rock	$\surd$		$\surd$		aggression, anger, rage, irritation
	Hip	$\surd$		$\surd$	$\surd$	dancing, anger, rebellion

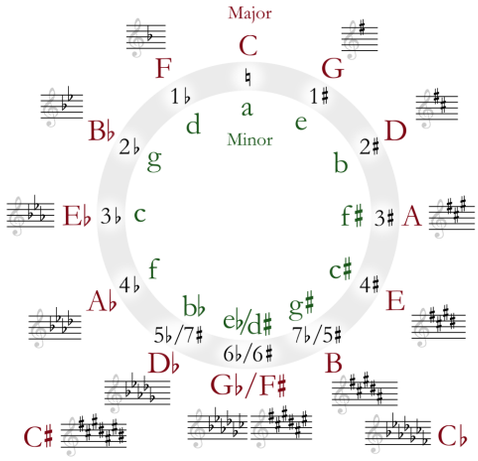
Composers use different musical elements and styles to communicate emotions to audiences. In this section, we introduce how music affects audiences' emotion from the perspectives of musical elements and music styles.

6.2.1 Dimension-related Domain Knowledge

Rhythm, tonality, timbre, dynamics, and pitch, are the five main musical dimensions usually designed by composers to communicate music emotions to listeners [Zbikowski and Zbikowski, 2002][Lartillot, 2011]. In this paper, we adopt the MIR toolbox [Lartillot, 2011] to extract these musical element features.

6.2.1.1 Rhythm

Tempo is one of the essential elements of music representing the speed or pace of a given piece. Composers usually use it to affect the mood of audiences [Fernández-Sotos et al., 2016; Gabrielsson and Lindström, 2010; Gomez and Danuser, 2007; Husain et al., 2002]. Generally, tempo expresses the rhythm and fluency of the music[Lartillot, 2011] and it is measured by the number of beats per minute. Commonly, tempo is categorized into Larghissimo (24 bpm and under), Grave (25-45 bpm), Lento (45-60 bpm), Larghetto (60-66 bpm), Adagio (66-76 bpm), Andante (76-108 bpm), Moderato (98-112 bpm), Allegro moderato (112-120 bpm), Allegro (120-156 bpm), Vivace (156-176 bpm), Presto (168-200 bpm) and Prestissimo (200bpm and over). By



No.	Flats		Sharps	
	Major	minor	Major	minor
0	C	a	C	a
1	F	d	G	e
2	B♭	g	D	b
3	E♭	c	A	f♯
4	A♭	f	E	c♯
5	D♭	b♭	B	g♯
6	G♭	e♭	F♯	d♯
7	C♭	a♭	C♯	a♯
8	F♭	d♭	G♯	e♯

Figure 6.2: The major mode and minor mode of music corresponding to stave and musical mark.

changing the tempo, the composers can evoke various emotions from the listeners via their music.

6.2.1.2 Tonality

Tonality is the arrangement of pitch and/or chords of a musical work in a hierarchy of perceived relations, stabilities, attractions and directionality. Composers use the mode of music to control tonality due to a system of musical tonality involved with mode [Miller, 2005]. Mode comprises major mode and minor mode in which the notes in the corresponding major scale or minor scale are considered diatonic notes, and other notes are considered chromatic notes [Jones, 1974]. Figure 6.2 shows the relative major modes and minor modes. From Figure 6.2, we can find the music is major or minor according to the mark of stave. For example, in Figure 7.1, we find a "#" in the stave and thus we conclude this musical mode is G major.

6.2.1.3 Timbre

Term "brightness" is used in the discussions on musical timbre which is a powerful component in constructing the sound quality of a musical piece from psychoacoustics[Wessel, 1979]. Composers usually uses different musical instruments, such as percussion instruments, wind instruments, and string instruments to express musical brightness to distinguish all kinds of sound production. In this way, composers can affect the emotions of their listeners, for example, the audience can feel joyful, angry or thrilled when the music has a bright tone; or they can feel depressed or sad when the music has a dull sound[Gabrielsson and Lindström, 2010].

Table 6.2: Scale of dynamic markings.

Name	Letters	Level
fortississimo	**fff**	very very loud
fortissimo	**ff**	very loud
forte	**f**	loud
mezzo-forte	**mf**	average
mezzo-piano	**mp**	
piano	**p**	soft
painissimo	**pp**	very soft
pianississimo	**ppp**	very very soft

6.2.1.4 Dynamics

Dynamics, which is the variation in loudness between notes or phrases, is one of the expressive elements of music. Used effectively, dynamics help musicians sustain variety and interest in a musical performance and communicate a particular emotional state or feeling[Gabrielsson and Lindström, 2010]. Dynamic indications have been marked in the music scores shown in Table 6.2.

6.2.1.5 Pitch

Pitch is usually used to judge the sounds as higher and lower in relation to musical melody [Plack et al., 2006]. Studies have shown that high-pitched music is associated with high arousal or valence such as tension, excitement, joy, happiness etc. while low-pitched music tends to be associated with low arousal or valence such as serenity, gratefulness and anger etc. [Ilie and Thompson, 2006; Scherer and Oshinsky, 1977][Hevner, 1937].

6.2.2 Style-related Domain Knowledge

Music can be described in terms of many genres and styles, which are useful for organizing similar musical characteristics from songs, albums and artists into broader groups. Currently, vast works have explored the topic of music genre classification using machine learning techniques and deep learning methods. For example, Oramas *et al.* [Oramas et al., 2018] proposed to learn intermediate representations of deep neural networks from audio tracks, text reviews etc. and combine these representations for music genre classification. Bahuleyan *et al.* [Bahuleyan, 2018] used a CNN model trained end-to-end to predict the genre label of an audio signal. However, the domain knowledge of different genres and music emotion has not been thoroughly exploited. In this section, we summarize the different music genres and

styles, and explore the domain knowledge from musicology and psychology. There is a list of music styles including pop, soul blues, classical, country, electronic, folk, international, jazz, rock, hip etc [1]. By varying their arrangements of these music styles, composers can evoke different emotions from their listeners. These styles are described as follows:

- **Pop** - a genre of popular music comprising short to medium-length songs written in a basic format (often the verse-chorus structure), as well as the common use of repeated choruses, melodic tunes and hooks [Frith et al., 2001].
- **Blues** - a music genre and musical form which incorporate spirituals, work songs, field hollers, shouts, chants, and rhymed simple narrative ballads. [Erlewine, 1999].
- **Classical** - art music produced or rooted in the traditions of Western culture, including both liturgical (religious) and secular music [Heyman, 2019].
- **Country** - a genre of popular music that takes its roots from genres such as blues and old-time music, and various types of American folk music [Peterson, 1978].
- **Electronic** - music that employs electronic musical instruments, digital instruments and circuitry-based music technology [Holmes and Holmes, 2002].
- **Folk** - as music transmitted orally, music with unknown composers, or music performed by custom over a long period of time [Bohlman, 1988].
- **International** - a Western musical category encompassing many different styles of music from other parts of the globe [Best et al., 2014].
- **Jazz** - characterized by swing and blue notes, call and response vocals, polyrhythms and improvisation [Starr and Waterman, 2008].
- **Rock** - a broad genre of popular music that originated as "rock and roll" [Studwell and Lonergan, 1999].
- **Hip** - stylized rhythmic music that commonly accompanies rapping, a rhythmic and rhyming speech that is chanted[Lightstone, 2012].

Previous research has found that jazz is associated with longing, amazement, spirituality and peacefulness [Susino and Schubert, 2019]; classical music is related to

[1] https://www.allmusic.com/genres

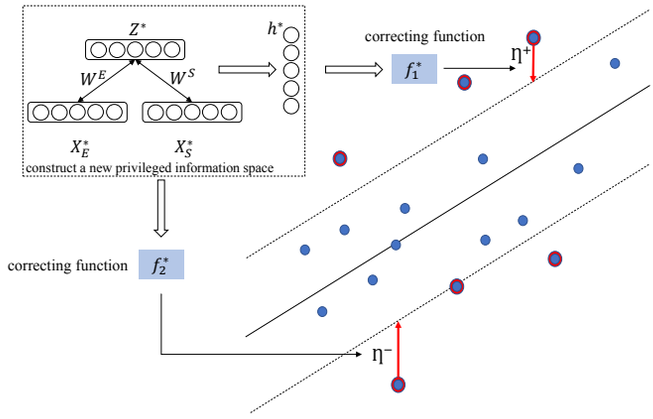


Figure 6.3: The framework of the proposed method.

calmness, relaxation, peace and joy [Susino and Schubert, 2019]; rock music is associated with aggression, anger, rage, irritation and revolt [Susino and Schubert, 2019]. Blues and jazz are also significantly related to the reported use of music for positive and negative emotion regulation strategies [Cook et al., 2019]. Energetic and rhythmic music, such as hip and electronic, are associated with the personality traits extraversion and agreeableness [Rentfrow and Gosling, 2003]. Country and pop music are positively associated with increased arousal and folk is associated only with negative mood management [Cook et al., 2019].

Table 6.1 summarizes the effects and emotion description of musical elements and music styles on the arousal and valence space. From Table 6.1, we obtain the relations between musical elements and A-V space as well as the relations between music styles and A-V space. This indicates that musical elements representing five musical dimensions and music styles representing music genres are informative and play an important role in music emotion recognition. Hence, in this paper, we regard them as privileged information [Vapnik and Vashist, 2009] to assist and enhance our task.

6.3 Methodology

In this section, we introduce the method to utilize the summary of domain knowledge for music emotion recognition. The subjects were asked to rate the valence and arousal values that best describe their general emotion perception of each clip.

6.3.1 Problem Statement

In this section, we define the problem and state the assumptions we used. The framework of the proposed method is summarized in Figure 6.3. As shown in Figure 6.3, there are two components in the proposed method: the RBM model and the privileged information-based support vector regression (SVR+). The RBM model captures the intrinsic dependencies among the visible variables represented by musical elements and music styles, then generates the new privileged information features. The SVR+ utilizes the generated privileged information during training to estimate the slack variables by two linear correcting functions. During testing, the music emotion of an unknown music piece or song is predicted by the learned music emotion estimator.

Denote a set of quadruple $S = \{\mathbf{X}, \mathbf{X_E^*}, \mathbf{X_S^*}, \mathbf{Y}\}$, where $\mathbf{X} \in \mathbb{R}^{N*d}$ is a feature matrix, $\mathbf{X_E^*} \in \mathbb{R}^{N*d_E^*}$ and $\mathbf{X_S^*} \in \mathbb{R}^{N*d_S^*}$ are the privileged information matrices of musical elements and music styles respectively. N is the number of training instances. d, d_E^* and d_S^* are the dimensions of feature, privileged information of musical elements, privileged information of music styles respectively. $\mathbf{Y} \in \mathbb{R}^N$ represents continuous valence and arousal values.

Our goals is to learn a regression function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ with the help of two kinds of privileged information X_E^* and X_S^*. During evaluation, given a song, we perform emotion estimation as:

$$\hat{y} = f(x; \theta). \tag{6.1}$$

where θ is the parameter of the function f and $\hat{y}$ is the prediction of emotion labels. $\mathbf{X_E^*}$ and $\mathbf{X_S^*}$ are not available in the evaluation phase.

6.3.2 Modeling Relations between Musical Elements and Music Styles by RBM

The dependencies among the visible variables can be captured by RBM which connects all the visible units through the latent units. Therefore, we propose to use RBM to model the natural connection between musical elements and music styles.

We divide the visible nodes into two sections, one section is the features of the musical elements and the other is the features of the music styles. On the one hand, the features of the musical elements are continuous, so we use the Gaussian units for these this visible nodes. On the other hand, the features of the music styles and hidden nodes are discrete, so we consider them to be binary stochastic units. Hence, we use musical elements $\mathbf{X_E^*}$ as continuous visible units and music styles $\mathbf{X_S^*}$ as binary

visible units. d_E^* and d_S^* are their dimensions respectively. $\mathbf{z}^* \in \{0,1\}^{n_z}$ represents binary hidden units, where n_z means the number of hidden nodes. The energy of the state of our Gaussian-Bernoulli RBM [Hinton and Salakhutdinov, 2006] is shown as follows:

$$\begin{aligned} E(X_E^*, X_S^*, z^* \mid \theta) &= \sum_{i=1}^{d_E^*} \frac{(x_{Ei}^* - \hat{b}_{Ei})}{2\sigma_{Ei}^2} + \sum_{i=1}^{d_S^*} \hat{b}_{Si} x_{Si}^* \\ &- \sum_{i=1}^{d_E^*} \sum_{j=1}^{n_z} \frac{x_{Ei}^*}{\sigma_{Ei}} W_{ij}^E z_j - \sum_{i=1}^{d_S^*} \sum_{j=1}^{n_z} x_{Si}^* W_{ij}^S z_j - \sum_{j=1}^{n_z} \hat{b}_{zj} z_j, \end{aligned} \tag{6.2}$$

where $\hat{\mathbf{b}} = \{\hat{\mathbf{b}}_E, \hat{\mathbf{b}}_S, \hat{\mathbf{b}}_z\}$ is the combined biases of visible units with musical elements, visible units with music styles and hidden units. σ represents a Gaussian distribution with the variance, $\theta = \{\hat{\mathbf{b}}, \sigma^2, \mathbf{W^E}, \mathbf{W^S}\}$ are the parameters of our model, where $\mathbf{W^E}$ and $\mathbf{W^S}$ are the weights between element-related, style-related visible units and hidden units. The joint distribution over visible units is defined by

$$P(X_E^*, X_S^* \mid \theta) = \frac{1}{Z(\theta)} \sum_{z^*} \exp(-E(X_E^*, X_S^*, z^* \mid \theta)), \tag{6.3}$$

where $Z(\theta) = \int_{x_E^*} \sum_{d_S^*} \sum_{n_z} \exp(-E(X_E^*, X_S^*, z^* \mid \theta)) dx_E^*$.

Then, the conditional distributions are given in Eq. 6.4, Eq. 6.5 and Eq. 6.6.

$$\begin{aligned} &P(X_E^* \mid z^*; \theta) = \prod_{i=1}^{d_E^*} p(x_{Ei}^* \mid z^*), \\ &\text{with } x_{Ei}^* \mid z^* \sim N(\sigma_{Ei} \sum_{i=1}^{n_z} W_{ij}^E \cdot z_j^* + \hat{b}_{Ei}, \sigma_{Ei}^2) \end{aligned} \tag{6.4}$$

$$\begin{aligned} &P(X_S^* \mid z^*; \theta) = \prod_{i=1}^{d_S^*} p(x_{Si}^* \mid z^*), \\ &\text{with } p(x_{Si}^* \mid z^*) = g(\sum_{i=1}^{n_z} W_{ij}^S \cdot z_j^* + \hat{b}_{Si}) \end{aligned} \tag{6.5}$$

$$\begin{aligned} &P(z^* \mid X_E^*, X_E^*; \theta) = \prod_{i=1}^{n_z} p(z_i^* \mid X_E^*, X_S^*), \\ &\text{with } p(z_i^* \mid X_E^*, X_S^*) = g(\sum_{i=1}^{d_E^*} W_{ij}^E \frac{x_{Ei}^*}{\sigma_{Ei}} + \sum_{i=1}^{d_S^*} W_{ij}^S x_{Si}^* + \hat{b}_{zj}) \end{aligned} \tag{6.6}$$

where $g(x)$ represents the sigmoid function $1/(1+\exp(-x))$. Given a set of obser-

vations $\{x^*_{Em}, x^*_{Sm}\}_{m=1}^{M}$, the derivative of the log-likelihood with respect to W^E and W^S are shown in Eq. 6.7 and Eq. 6.8 respectively.

$$\frac{1}{M}\sum_{m=1}^{M}\frac{\partial logP(x^*_{Em}, x^*_{Sm}; \theta)}{\partial W^E_{ij}} = E_{P_{data}}[\frac{x^*_{Ei}}{\sigma_{Ei}} z^*_j] - E_{P_{model}}[\frac{x^*_{Ei}}{\sigma_{Ei}} z^*_j] \tag{6.7}$$

$$\frac{1}{M}\sum_{m=1}^{M}\frac{\partial logP(x^*_{Em}, x^*_{Sm}; \theta)}{\partial W^S_{ij}} = E_{P_{data}}[x^*_{Si} z^*_j] - E_{P_{model}}[x^*_{Si} z^*_j]. \tag{6.8}$$

The Contrastive Divergence (CD) learning algorithm are used to learn the parameters θ. The details are in [Hinton, 2012].

After learning the model parameters, we sample the hidden layer and the visible layer by updating the music style using Eq. 6.9 and the hidden layer using Eq. 6.10 with Gibbs sampling.

$$\mu_j \leftarrow g(\sum_{i=1}^{n_z} W^S_{ij} \cdot h^*_i + \hat{b}_{Sj}) \tag{6.9}$$

$$h^*_i \leftarrow g(\sum_{j=1}^{d^*_E} W^E_{ij} \frac{x^*_{Ej}}{\sigma_{Ej}} + \sum_{j=1}^{d^*_S} W^S_{ij} \mu_j + \hat{b}_{zi}). \tag{6.10}$$

After μ_j and h^*_i are converged, we use the h^*_i as the new privileged information for music emotion recognition.

6.3.3 Proposed LUPI for Music Emotion Recognition

The new privileged information representation $h^*_i \in \mathbf{h}^* \in \mathbb{R}^{N * d^*_h}$ from integrating musical elements and styles is used to refine model estimation during training.

Firstly, we briefly review Support Vector Regression (SVR). One variant called ϵ-insensitive SVR is given by

$$\begin{aligned} \min_{w,\eta^+,\eta^-} \ & \frac{1}{2}||\mathbf{w}||^2 + \gamma \sum_i (\eta^+_i + \eta^-_i) \\ \text{s.t. } & w^T\phi(X_i) + b - Y_i \leq \epsilon + \eta^+_i \\ & Y_i - w^T\phi(X_i) - b \leq \epsilon + \eta^-_i \\ & \eta^+_i, \eta^-_i \geq 0, \ \forall i \end{aligned} \tag{6.11}$$

where ϵ is a constant that defines the maximum deviation allowed for a prediction to be considered as correct. ϕ is a kernel function that maps X_i into another feature space. The performance of SVR relies heavily on the choice of kernel function. γ is a trade-off between the model complexity and model loss.

Then, we consider three sets of linear functions as follows:

$$\begin{aligned} f(X_i) &= (w, \phi(X_i)) + b, \\ f_1^*(X_i^*) &= (w_1^*, \phi(X_i^*)) + b_1^*, \\ f_2^*(X_i^*) &= (w_2^*, \phi(X_i^*)) + b_2^* \end{aligned} \tag{6.12}$$

where w, w_1^* and w_2^* are the weight coefficients of the general feature regressor and privileged information regressor, respectively, b, b_1^* and b_2^* are their bias. $\phi(X_i)$ and $\phi(X_i^*)$ project X_i and X_i^* into the kernel space, and (,) indicates the matrix product of the two terms in the bracket.

We define our slack variable $\eta^+ = (w_1^*, \phi(h_i^*)) + b_1^*$ and $\eta^- = (w_2^*, \phi(h_i^*)) + b_2^*$ respectively [Vapnik and Vashist, 2009].

Therefore, the objective function of the proposed method is as follows:

$$\begin{aligned} \min_{w, w_1^*, w_2^*} & \sum_{i=1}^{N} \frac{1}{2}(||w_i||^2 + ||w_{1i}^*||^2 + ||w_{2i}^*||^2) \\ & + \gamma \sum_{i=1}^{N}[(w_1^*, \phi(h_i^*)) + b_1^*] + \gamma \sum_{i=1}^{N}[(w_2^*, \phi(h_i^*)) + b_2^*] \\ \text{s.t. } & Y_i - w_i^T \phi(X_i) - b \leq \epsilon + (w_{1i}^*, \phi(h_i^*)) + b_{1i}^* \\ & w_i^T \phi(X_i) + b - Y_i \leq \epsilon + (w_{2i}^*, \phi(h_i^*)) + b_{2i}^* \\ & [(w_{1i}^*, \phi(h_i^*)) + b_{1i}^*] \geq 0 \\ & [(w_{2i}^*, \phi(h_i^*)) + b_{2i}^*] \geq 0, \ \forall i \end{aligned} \tag{6.13}$$

In order to solve this convex optimization problem, we apply the Lagrangian multiplier method [Everett III, 1963] to simplify the formulation and improve com-

putational efficiency. The Lagrangian function is shown as follows:

$$\begin{aligned} L = & \sum_{i=1}^{N} \frac{1}{2}(||w_i||^2 + ||w_{1i}^*||^2 + ||w_{2i}^*||^2) \\ & + \gamma \sum_{i=1}^{N}[(w_{1i}^*, \phi(h_i^*)) + b_{1i}^*] + \gamma \sum_{i=1}^{N}[(w_{2i}^*, \phi(h_i^*)) + b_{2i}^*] \\ & + \sum_{i=1}^{N} \alpha_i(Y_i - f(X_i) - \epsilon - f_1^*(h_i^*)) + \sum_{i=1}^{N} \beta_i(0 - f_1^*(h_i^*)) \\ & + \sum_{i=1}^{N} \alpha_i^*(f(X_i) - Y_i - \epsilon - f_2^*(h_i^*)) + \sum_{i=1}^{N} \beta_i^*(0 - f_2^*(h_i^*)) \end{aligned} \tag{6.14}$$

where α, α^*, β, β^* are Lagrangian multipliers. The gradient of function Eq. 6.14 to the weight and bias can be computed by Eq. 6.15.

$$\begin{aligned} \frac{\partial L}{\partial w_i} &= w_i - \sum_{i=1}^{N} \alpha_i \phi(X_i) + \sum_{i=1}^{N} \alpha_i^* \phi(X_i) \\ \frac{\partial L}{\partial w_{1i}^*} &= w_{1i}^* + \gamma \sum_{i=1}^{N} \phi(X_i) - \sum_{i=1}^{N} \alpha_i \phi(h_i^*) - \sum_{i=1}^{N} \beta \phi(h_i^*) \\ \frac{\partial L}{\partial w_{2i}^*} &= w_{2i}^* + \gamma \sum_{i=1}^{N} \phi(X_i) - \sum_{i=1}^{N} \alpha_i^* \phi(h_i^*) - \sum_{i=1}^{N} \beta^* \phi(h_i^*) \\ \frac{\partial L}{\partial b_i} &= \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) \\ \frac{\partial L}{\partial b_{1i}^*} &= \sum_{i=1}^{N} (\gamma - \alpha_i - \beta_i) \\ \frac{\partial L}{\partial b_{2i}^*} &= \sum_{i=1}^{N} (\gamma - \alpha_i^* - \beta_i^*) \end{aligned} \tag{6.15}$$

Setting Eq. 6.15 to zero we get the following equation:

$$w_i = \sum_{i=1}^{N}(\alpha_i^* - \alpha)\phi(X_i) \tag{6.16a}$$

$$w_{1i}^* = \sum_{i=1}^{N}(\alpha_i + \beta_i - \gamma)\phi(X_i) \tag{6.16b}$$

$$w_{2i}^* = \sum_{i=1}^{N}(\alpha_i^* + \beta_i^* - \gamma)\phi(X_i) \tag{6.16c}$$

$$\sum_{i=1}^{N}(\alpha_i^* - \alpha_i) = 0 \tag{6.16d}$$

$$\gamma = \alpha_i + \beta_i \tag{6.16e}$$

$$\gamma = \alpha_i^* + \beta_i^* \tag{6.16f}$$

We get the dual problem to this object as follows:

$$\begin{aligned}
\min_{\alpha,\alpha^*,\beta,\beta^*} & \sum_{i=1}^{N}\gamma_i(\alpha_i - \alpha_i^*) + \epsilon(\alpha_i + \alpha_i^*) \\
& + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(X_i, X_j) \\
& + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^* + \beta_i^* - \gamma)(\alpha_j^* + \beta_j^* - \gamma)K^*(h_i^*, h_j^*) \\
& + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i + \beta_i - \gamma)(\alpha_j + \beta_j - \gamma)K^*(h_i^*, h_j^*) \\
\text{s.t. } & \sum_{i=1}^{N}(\alpha_i^* - \alpha_i) = 0 \\
& \sum_{i=1}^{N}(\alpha_i^* + \beta_i^* - \gamma) = 0 \\
& \sum_{i=1}^{N}(\alpha_i + \beta_i - \gamma) = 0 \\
& \alpha_i^*, \alpha_i, \beta_i^*, \beta_i \geq 0 \; i = 1, ..., N.
\end{aligned} \tag{6.17}$$

where $K(\cdot,\cdot)$ and $K^*(\cdot,\cdot)$ are kernels that define the inner products for $\phi(X)$ and $\phi(X^*)$ respectively.

Then, according to the KKT condition, Eq. 6.16e and Eq. 6.16f, we get the follow-

ing:

$$\alpha_i(Y_i - f(X_i) - \epsilon - f_1^*(h_i^*)) = 0 \tag{6.18a}$$
$$\alpha_i^*(f(X_i) - Y_i - \epsilon - f_2^*(h_i^*)) = 0 \tag{6.18b}$$
$$\beta_i f_1^*(h_i^*) = (\gamma - \alpha_i) f_1^*(h_i^*) = 0 \tag{6.18c}$$
$$\beta_i^* f_2^*(h_i^*) = (\gamma - \alpha_i^*) f_1^*(h_i^*) = 0 \tag{6.18d}$$

The product of α_i and α_i^* in the dual problem is zero. i.e. $\alpha_i \alpha_i^* = 0$.

$\exists(x_k, y_k), k \in \mathbb{R}^N$, makes $f(x_k) - y_k - \epsilon - f_2^*(h_k^*) = 0$ or $y_k - f(x_k) - \epsilon - f_1^*(h_k^*) = 0$. i.e. the only one that's equal to 0. According to Eq. 6.18a and Eq. 6.18b, $\alpha_i = 0$ and $\alpha_i^* = 0$, one of them must be zero. q.e.d.

Define set $S_v = \{s_{vj} = (x_j, y_j) | j = 1, 2, ..., q\}$ is the set of support vectors and q as the number of support vectors. S_u and S_d are the sets of support vectors in the upper boundary and down boundary respectively. Then,

$$b = \frac{1}{|S_u| + |S_d|}[\sum_{s_{vj} \in S_u} (y_j - \epsilon - wx_j) + \sum_{s_{vj} \in S_d} (y_j + \epsilon - wx_j)] \tag{6.19}$$

$S_v = S_u \cup S_d$. For all support vectors (x_j, x_j):

$$\begin{aligned} &f(x_j) - y_j - \epsilon - f_2^*(h_j^*) = 0 \; or \\ &y_j - f(x_j) - \epsilon - f_1^*(h_j^*) = 0. \end{aligned} \tag{6.20}$$

According to Eq. 6.18c, we get:

$$\begin{aligned} &\text{if } 0 < \alpha_j < \gamma, \text{then } f_1^*(h_j^*) = 0; \\ &\text{if } 0 < \alpha_j^* < \gamma, \text{then } f_2^*(h_j^*) = 0 \end{aligned} \tag{6.21}$$

Because one of α_j and α_j^* equals 0, we get $f_1^*(h_j^*) = 0$ or $f_2^*(h_j^*) = 0$. Thus, we get: $f(x_j) = y_j + \epsilon$ or $f(x_j) = y_j - \epsilon$. The combination Eq. 6.16a and $b = f(x) - wx$, we get the result. q.e.d.

Finally, we get the parameters w and b respectively, the details as follows:

$$\begin{aligned} w &= \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) X_i, \\ b &= \frac{1}{|S_u| + |S_d|}[\sum_{s_{vj} \in S_u} (y_j - \epsilon - wx_j) + \sum_{s_{vj} \in S_d} (y_j + \epsilon - wx_j)], \end{aligned} \tag{6.22}$$

The two correcting functions are shown as follows:

$$\begin{aligned} f_1^*(h^*) &= (w_1^*, \phi(h^*)) + b_1^* = \sum_{i=1}^{N} (\alpha_i + \beta_i - \gamma) K^*(h_i^*, h^*) + b_1^*, \\ f_2^*(h^*) &= (w_2^*, \phi(h^*)) + b_2^* = \sum_{i=1}^{N} (\alpha_i^* + \beta_i^* - \gamma) K^*(h_i^*, h^*) + b_2^* \end{aligned} \tag{6.23}$$

Hence, the decision function for solving this problem is given as follows:

$$f(X) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) K(X_i, X) + b. \tag{6.24}$$

6.3.4 Optimization

The learning algorithm of the proposed method is divided into two procedures: one is using RBM to generate new privileged information via modeling the intrinsic relations between musical elements and music styles; the other is formulating SVR+ with new privileged information as a constrained optimization problem. We outline the learning procedure of the proposed method in Algorithm 3.

6.4 Experiments

To demonstrate the effectiveness of the proposed method, we conduct experiments on two benchmark databases: MediaEval [2] and the All Music Guide 1608 database (AMG 1608) [Chen et al., 2015].

6.4.1 Experimental Conditions

6.4.1.1 Datasets

The MediaEval database is split into a training set and a testing set. The training set contains 430 music clips with each clip being 45 seconds in length, and the testing set contains 58 complete music pieces with an average of 234 seconds per clip. We use 260 low-level feature sets provided by [Aljanaki et al., 2015], which are extracted using openSMILE toolbox [Eyben et al., 2010]. The 260-dimensional feature set, including RMS ennergy, MFCC, ZCR and F0 etc. represents the music from 65-dimensional mean deviations, 65-dimensional standard deviations, and their first-order derivatives from acoustic descriptors. A two-dimensional space of continuous real values

[2] http://www.multimediaeval.org/mediaeval2015/emotioninmusic2015/

Algorithm 3 Training algorithm of the proposed model

Require:
training samples$\{\mathbf{X}, \mathbf{X}^*_{\mathbf{E}}, \mathbf{X}^*_{\mathbf{S}}, \mathbf{Y}\}$, coefficient γ
the parameters $\{\mathbf{W^E}, \mathbf{W^S}, \hat{\mathbf{b}}\}$
Ensure: Model parameters w, b
Randomly initialize h^*_i, μ_j;
repeat
for not converge **do**
Update the states of each hidden units.

$$h^*_i \leftarrow g(\sum_{j=1}^{d^*_E} W^E_{ij} \frac{x^*_{Ej}}{\sigma_{Ej}} + \sum_{j=1}^{d^*_S} W^S_{ij} \mu_j + \hat{b}_{zi})$$

Update the states of visible units.

$$\mu_j \leftarrow g(\sum_{i=1}^{n_z} W^S_{ij} \cdot h^*_i + \hat{b}_{Sj})$$

until Converges
Fetch new privileged information h^*_i
Project $\mathbf{X}$ and $\mathbf{h}^*$ into kernel space
Using conditional gradient method to solve Eq. 6.17
Fetch the Lagrangian multipliers $\alpha^*_i, \alpha_i, \beta^*_i, \beta_i$
return w, b

between [–1, 1] was provided for annotators to indicate the valence and arousal values at the same time. The official split of the MediaEval database does not include a validation set, and thus we have to construct the validation set by ourselves. We randomly sampled 58 music clips from the training set as the validation set, and left the remaining 372 music clips for training.

The AMG 1608 database consists of 1608 preview clips of western songs, collected from a popular music stream service named 7digit. Each preview clips is 30-second long. We use the public feature set provided by [Chen et al., 2015], including MFCC, Tonal, Spectral and Temporal. Similar to MediaEval, the valence and arousal values, which are real values ranging in between [–1, 1]. For the experiments, we adopt four-fold cross-validation on the database.

6.4.1.2 Privileged Information

We extract tempo, mode, brightness, loudness and pitch with the MIR toolbox[Lartillot, 2011] on these two databases. The music styles are provided in MediaEval, but not in the AMG 1608 database. Hence, we manually annotated the music styles of all the

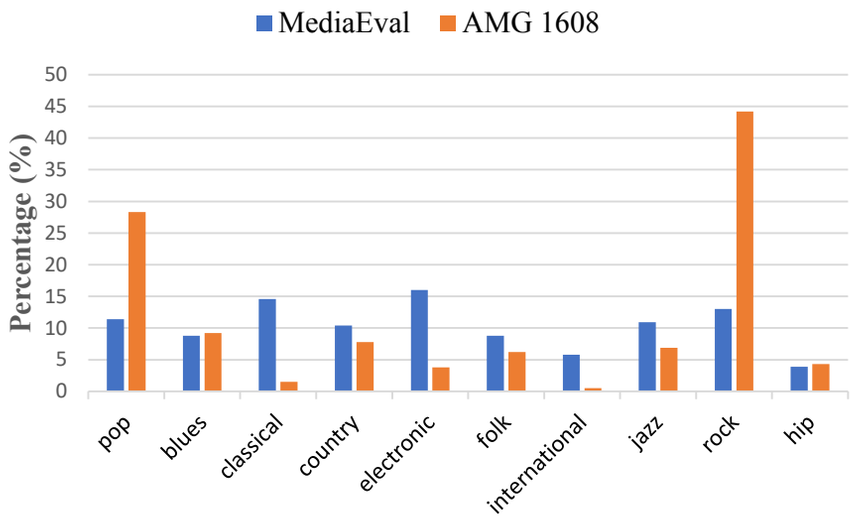


Figure 6.4: The distributions of all music styles on both MediaEval and AMG 1608 databases. The X-axis is the different music styles and the Y-axis is their percentage in the corresponding databases.

songs in the AMG 1608 database as shown in Figure 6.4. The statistics of the music styles in these two databases are different. Songs on the MediaEval database have a similar proportion of music styles. However, on the AMG 1608 database, most of the songs' styles are rock and pop. This information is considered for analyzing the impact on different distributions of music styles.

6.4.1.3 Evaluation Details

The evaluation metrics we utilize to evaluate the effectiveness of the proposed method are root mean square error (RMSE) and Pearson's correlation (R). In terms of RMSE, the lower the better and for R, the larger the better. We first randomly initialize the weights with small value. Then we use the grid search to conduct parameter selection. For each method, we select the parameters on the validation set in order to minimize the objective cost.

6.4.1.4 Baselines

We choose the following eight popular methods for comparison on the MediaEval 2015 database.

- Multiple linear regression (MLR) a common linear regression method, which is used as a baseline in current works in music emotion recognition [Aljanaki et al., 2015].

- recurrent neural networks (RNN) [Pellegrini and Barriere, 2015]: a method that has sequence modeling capabilities to predict two real-valued emotion dimensions, i.e. arousal space and valence space, in a time-continuous estimation of emotion in music.

- SVR [Chmulik et al., 2015]: the proposed method is based on Support Vector Regression with Radial Basis kernel function for music emotion recognition.

- Double-scale support vector regression (DS+SVR)[Xianyu et al., 2016] recognize time-continuous emotion in music, which decouple the two scales of dynamics, i.e. the glob-scale dynamics and the local-scale dynamics for obtaining the recognizing result.

- deep long-short term memory recurrent neural networks (LSTM-RNN) utilize acoustic and psychoacoustic features for dynamic arousal and valence prediction [Coutinho et al., 2015].

- Deep BLSTM (DBLSTM) utilize the advantage and ability of model's sequence information to capture the high context correlation among the music feature sequence for predicting the valence and arousal values in music emotion recognition [Li et al., 2016b].

- deep convolutional neural networks (DNN) is based on convolutional and recurrent neural networks where this model uses bidirectional gated recurrent unit to stack convolution layer to improve the performance of emotion recognition [Orjesek et al., 2019].

- DKLR model captures the relations between common high-level music features and emotion as constraints into the proposed objective function to enhance the regression model for music emotion recognition [Shu and Xu, 2019].

On another database of music emotion recognition, AMG 1608, we compare our method with five popular methods. Two of them (MLR and DKLR) are the same as MediaEval 2015, and the other three works are detailed as follows:

- The Acoustic Emotion Gaussians (AEG) model [Wang et al., 2016] uses probability distributions to account for the subjectivity of emotion prediction on the arousal and valence space.

- CDCC [Hu and Yang, 2017] focus on cross-dataset generalizability on the related fifteen acoustic features. This regression is built to predict the values of valence and arousal for examining the mood regression performance.

- The twin Gaussian process (TGP) [Chapaneri and Jayaswal, 2017] is used for structured regression to model the arousal and valence dimensions of mood.

6.4.2 Comparison Results

Table 6.3: Comparison with the related works on the MediaEval 2015 database.

Model	Arousal		Valence	
	RMSE	R	RMSE	R
MLR [Aljanaki et al., 2015]	0.27	0.36	0.37	0.01
RNN [Pellegrini and Barriere, 2015]	0.247	0.588	0.365	0.029
SVR [Chmulik et al., 2015]	0.255	0.510	0.366	0.022
DS+SVR [Xianyu et al., 2016]	0.234	0.61	0.366	0.02
LSTM+RNN [Coutinho et al., 2015]	0.242	0.611	0.373	0.0004
DBLSTM [Li et al., 2016b]	0.239	-	0.318	-
DNN [Orjesek et al., 2019]	0.214	-	0.240	-
DKLR [Shu and Xu, 2019]	0.234	0.597	0.318	0.044
Ours	**0.194**	**0.633**	**0.201**	**0.132**

Table 6.4: Comparison with the related works on the AMG 1608 database.

Model	Arousal		Valence	
	RMSE	R	RMSE	R
MLR [Aljanaki et al., 2015]	0.288	0.806	0.288	0.346
AEG [Wang et al., 2016]	0.287	0.809	0.287	0.400
CDCC [Hu and Yang, 2017]	-	0.854	-	0.435
DKLR [Shu and Xu, 2019]	0.240	0.817	0.254	0.374
Regression with TGP [Chapaneri and Jayaswal, 2017]	0.203	0.808	0.236	0.661
Ours	**0.196**	**0.891**	**0.228**	**0.677**

The comparison results on the MediaEval 2015 database are shown in Table 6.3. All the above methods focus on different classifiers or models to boost the emotion recognition performance, but they ignore the informative domain knowledge on music emotion. In this paper, we propose a method to improve music emotion recognition by exploiting the domain knowledge of musical elements and music genres from musicology and psychology and use an RBM model to thoroughly model their dependency relations.

Table 6.4 shows the experimental results of all comparison methods on the AMG 1608 database. Compared to these works, we utilize the simplest features but achieve the best performance, which indicates the vital role of privileged information in our methods. The proposed method uses the privileged information of RBM fusion be-

tween the main musical elements and music styles, resulting in a superior performance compared to state-of-the-art methods.

6.4.3 Analysis of Fusion Methods

Table 6.5: Emotion recognition on MediaEval 2015 database.

	Arousal						Valence					
	SVR feat-level fusion	SVR deci-level fusion	SVR+ elements	SVR+ styles	SVR+ fusion	Our model	SVR feat-level fusion	SVR deci-level fusion	SVR+ elements	SVR+ styles	SVR+ fusion	Our model
RMSE	0.280	0.273	0.224	0.217	0.215	**0.194**	0.382	0.364	0.264	0.251	0.233	**0.201**
R	0.357	0.368	0.586	0.588	0.602	**0.633**	0.008	0.014	0.116	0.108	0.121	**0.132**

Table 6.6: Emotion recognition on AMG 1608 database.

	Arousal						Valence					
	SVR feat-level fusion	SVR deci-level fusion	SVR+ elements	SVR+ styles	SVR+ fusion	Our model	SVR feat-level fusion	SVR deci-level fusion	SVR+ elements	SVR+ styles	SVR+ fusion	Our model
RMSE	0.292	0.284	0.217	0.224	0.211	**0.196**	0.277	0.272	0.237	0.233	0.231	**0.228**
R	0.796	0.811	0.874	0.883	0.886	**0.891**	0.348	0.352	0.588	0.643	0.652	**0.677**

To investigate the effects of fusion methods of musical elements and music styles for the music emotion recognition, we conduct the following six experiments on both arousal and valence spaces on the two databases: feature-level fusion on SVR (**SVR feat-level fusion**); decision-level fusion on SVR (**SVR deci-level fusion**); musical element-related SVR+ (**SVR+ elements**); music style-related SVR+ (**SVR+ styles**); fusing musical elements and styles on SVR+ (**SVR+ fusion**); and **our model**. For feature-level fusion, the musical elements, music styles and acoustic, psychoacoustic features are concatenated as the input of SVR. For decision-level fusion, a weighted strategy is adopted to combine the estimation results of musical elements and styles from two SVRs. For musical element-related SVR+, music style-related SVR+ and fusion SVR+, the privileged information is musical elements, music styles and concatenation of them, respectively. For our method, the privileged information is estimated via new feature representation generated by RBM.

The experiment results of music emotion recognition assisted by musical elements and music styles are shown in Table 6.5 and Table 6.6. From these two tables, we make the following observations:

First, compared to the model using musical elements and music styles as features (**SVR feat-level fusion** and **SVR deci-level fusion**), the models using musical elements and music styles as privileged information integrated into the leaning process of the classifier achieve better performance. Specifically, the RMSE results of our model are 0.084/0.079 and 0.181/0.163 lower than that of **SVR feat-level fusion**/**SVR deci-level fusion**, and the R results of our model are 0.276/0.265 and 0.124/0.018 higher on the arousal and valence space on the MediaEval 2015 database.

Similar experiment results are achieved on the AMG 1608 database. This indicates that utilizing musical elements and music styles as privileged information are better than regarding musical elements and music styles as features.

Second, compared to the models using musical elements and using music styles as privileged information, the model which uses concatenated information by combining musical elements and music styles achieves better performance. Specifically, the RMSE and R of the **SVR+ fusion** model are 0.009/0.002 and 0.016/0.014 better than those of **SVR+ elements**/**SVR+ styles** for the arousal space, and 0.031/0.018 and 0.005/0.013 better for the valence space. Similar results are observed on the AMG 1608 database. The reason for this is that musical elements represent the main dimension and structure of music, while music styles represent the genres of music. They represent the song from different aspects, and their effects on music emotion recognition are complementary. Therefore, leveraging more informative privileged information results in superior recognition performance.

Third, our proposed method using RBM to capture the relation between musical elements and music styles and generate new privileged information into SVR achieve the best performance. Specifically, the RMSE and R are 0.021/0.031 better than the **SVR+ fusion** model on arousal and 0.032/0.011 on valence on the MediaEval 2015 database. Similar results are observed on the AMG 1608 database.

6.4.4 Comparison with Different Sorts of Privileged Information Features

Table 6.7: Exploring musical element-related privileged information for emotion recognition from songs on both the MediaEval 2015 and AMG 1608 databases. "-/-" denotes RMSE and R respectively, "total" represents the all musical elements.

Method	MediaEval		AMG 1608	
	arousal	valence	arousal	valence
none	0.286/0.345	0.388/0.006	0.297/0.788	0.283/0.342
tempo	0.246/0.567	0.292/0.102	0.238/0.843	0.256/0.548
mode	0.251/0.562	0.288/0.009	0.241/0.832	0.243/0.562
brightness	0.242/0.558	0.304/0.104	0.236/0.852	0.254/0.557
loudness	0.253/0.566	0.285/0.108	0.231/0.848	0.260/0.538
pitch	0.258/0.549	0.316/0.007	0.247/0.835	0.347/0.452
total	**0.224/0.586**	**0.264/0.116**	**0.217/0.874**	**0.237/0.588**

To investigate the effects of the combination of musical elements and music styles, we conduct three experiments considering the five dimensions of musical elements and the ten genres of music styles respectively on the two databases. The first is music emotion recognition without any musical element (**none**). The second is music

Table 6.8: Exploring music style-related privileged information for emotion recognition from songs on both the MediaEval and AMG 1608 databases. "-/-" denotes RMSE and R respectively, "total" represents all the music styles.

Method	MediaEval 2015		AMG 1608	
	arousal	valence	arousal	valence
none	0.284/0.349	0.393/0.006	0.298/0.788	0.286/0.343
pop	0.246/0.495	0.324/0.008	0.265/0.824	0.262/0.457
blues	0.245/0.512	0.325/0.012	0.258/0.838	0.268/0.432
classical	0.238/0.504	0.304/0.033	0.271/0.807	0.264/0.448
country	0.241/0.508	0.296/0.037	0.262/0.817	0.251/0.541
electronic	0.236/0.517	0.317/0.018	0.255/0.826	0.248/0.596
folk	0.253/0.498	0.306/0.018	0.258/0.823	0.257/0.567
intl	0.256/0.488	0.349/0.008	0.271/0.796	0.266/0.491
jazz	0.246/0.502	0.295/0.033	0.253/0.839	0.243/0.542
rock	0.231/0.524	0.288/0.052	0.250/0.842	0.241/0.598
hip	0.236/0.497	0.315/0.026	0.267/0.811	0.267/0.471
total	**0.217/0.588**	**0.251/0.108**	**0.224/0.883**	**0.233/0.643**

Table 6.9: The performance of the P-values of the proposed SVR+ with musical elements, music styles and privileged information of RBM compared with the method which ignores privileged information. "-/-" denotes P-values of RMSE and R respectively.

Method	MediaEval		AMG 1608	
	arousal	valence	arousal	valence
SVR+ elements	0.012/0.023	1.80E-03/2.34E-03	0.006/0.018	2.12E-03/1.43E-03
SVR+ styles	3.05E-03/1.14E-03	0.017/0.001	3.08E-03/1.62E-03	0.003/0.024
SVR+ RBM fusion (Our model)	1.63E-04/2.86E-04	1.71E-04/2.12E-04	3.24E-04/1.48E-04	1.26E-05/2.28E-04

emotion recognition with a single musical element, i.e. **tempo**, **mode**, **brightness**, **loudness**, **pitch**. The third is music emotion recognition with all musical elements (**total**). Similar experiments are conducted in relation to music styles. We show the experiment results for music emotion recognition on Table 6.7 and Table 6.8 and make the following observation.

First, compared to the model without any musical elements or music styles as privileged information (**none**), the model using single musical elements or music styles as privileged information performs better. This indicates that it is possible to perform emotion recognition from songs when using the five musical elements and the ten music styles. Thus, as expected, the model with the combined musical elements or music styles achieved better performance than that the model with no privileged information.

Second, each musical element or music style focuses on the songs from different aspects and make different contributions to emotion recognition. For example,

"pitch" describes the judgement of the frequency of the notes in the musical elements and has the worst RMSE and R in the arousal and valence space on the two databases. "Rock" describes a music style of the song centred on an amplified electric guitar and achieves the best results on the A-V space and the two databases. This indicates that the effects of musical elements and music styles on music emotion recognition are diversified and complementary.

6.4.5 Analysis the Effectiveness of Privileged Information

We conduct hypothesis tests to evaluate further whether the methods which use musical elements, music styles and RBM fusion as privileged information are significantly different from the method which ignores privileged information. To take into consideration that some of the experiment results may not follow a normal distribution, we adopt the two-sample non-parameter hypothesis, the Wilcoxon test, to evaluate the differences. Table 6.9 shows the experiment results of the P-values of the model which use privileged information and the model which does not on the AV space on the two databases. From Table 6.9, we observe that all P-values are less than 0.05, which indicates there is a significant difference between the proposed model and the model which ignores privileged information, i.e. musical elements and styles have a meaningful effect and help improve the performance of music emotion recognition.

6.5 Conclusion

6.5.1 Summary

In this chapter, we summarize the domain knowledge of musicology and exploit the relations among musical elements, music styles and musical emotion. We first propose the use of RBM model to capture their relations and construct a better privileged information representation. Then we use the generated new privileged information to estimate the slack variate of support vector regression involving the learning process and propose the novel SVR+ methods with musical elements and music styles of two kinds of musicology knowledge for music emotion recognition. Finally, we provide a detailed optimization algorithm to solve the proposed model. We conduct extensive experiments to demonstrate the superiority of the proposed SVR+ with privileged information method on the MediaEval 2015 and AMG 1608 database.

6.5.2 Future Work

In the future work, It would be better to combine the current audio feature MER with video affective content analysis, such as music in movies and music in videoland. The proposed method also can be extended to many fields via SVR+ or SVM+ model with domain knowledge as privileged information such as video sentiment analysis. We will explore these in future work.

Chapter 7

Learning Ordinal Privileged Information for Multiple Object Recognition

7.1 Introduction

The goal of multiple object recognition is to detect which instances the image contains, such as cars, people and dog.

Despite the progress witnessed in some learning algorithms, there still exist research gaps of how to exploit various forms of privileged information in efficient way. 1) *Ranking* form. Most existing models mainly handle privileged information in the *numeric* form, but ignores the *ranking (or ordinal)* form, which could be widely observed in various applications. To illustrate how the *ranking* form of privileged information can benefit regression tasks, we take an example shown in Figure 7.1. For six children with different heights shown in Fig.7.1(a), the task here is to estimate the height of the middle child ”4”, with the observed information from image that ”4” is higher than ”1, 2, 3” but lower than ”5, 6”, i.e. the privileged information in *ranking form*. In such case if we only know the height of ”3” and ”5” but ignores the ranking relation, we cannot estimate exactly the height of ”4”, which implies the important cues conveyed by the ranking should be fully leveraged. For Fig.7.1(b), the *rule of third* [Dhar et al., 2011], which means the attribute of objects closing to the four points of intersection can be considered as the privileged information for the aesthetic evaluation. Particularly, four main objects in the photo and the distance between these four objects and four points of intersection affect the photo aesthetics scores. Therefore, bringing in these additional ranking privileged information in training will benefit the learning, compared to using numeric privileged information standalone. 2) Efficiency. Currently only few studies used ranking privileged infor-

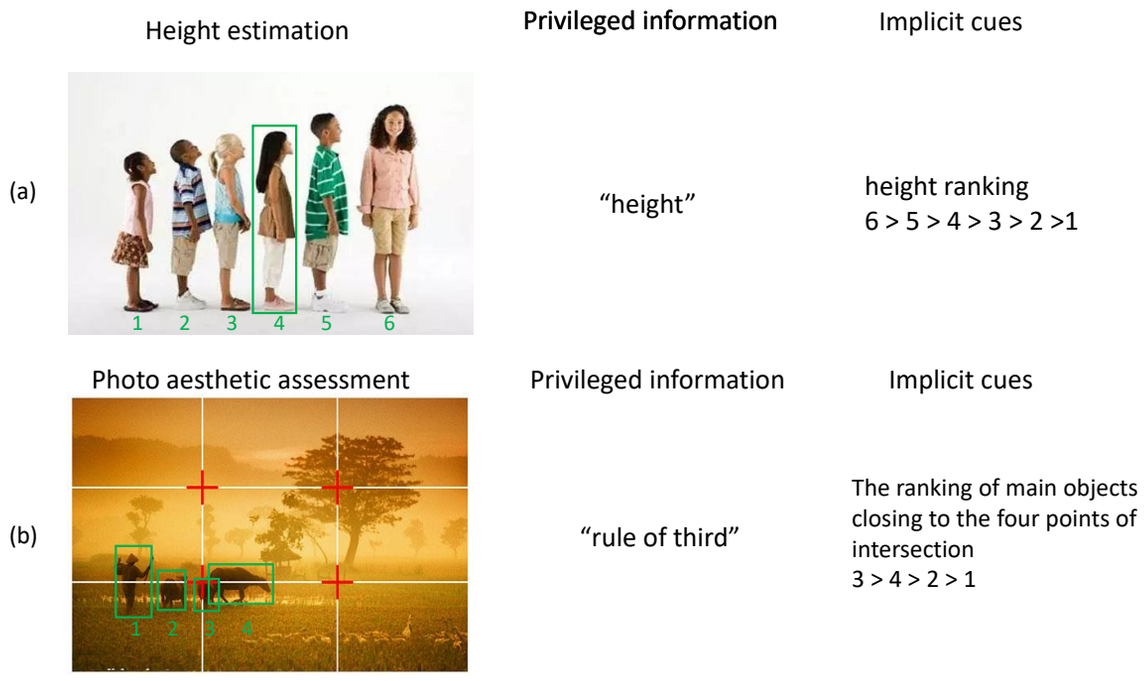


Figure 7.1: Examples of different privileged information forms

mation with suffering the efficiency issue, e.g. Sharmanska et al.[Sharmanska et al., 2013] proposed a two-step maximum-margin learning technique for privileged information. They first trained a ranking SVM on privileged information to obtain the ranking transfer function, then computed the margins to identify easy-to-separate and hard-to-separate example pairs by the ranking transfer function. Thus they used privileged information as ranking formula in constraint for twice, i.e. one for privileged information and another for original features, which increases computational complexity, leading to the inefficiency.

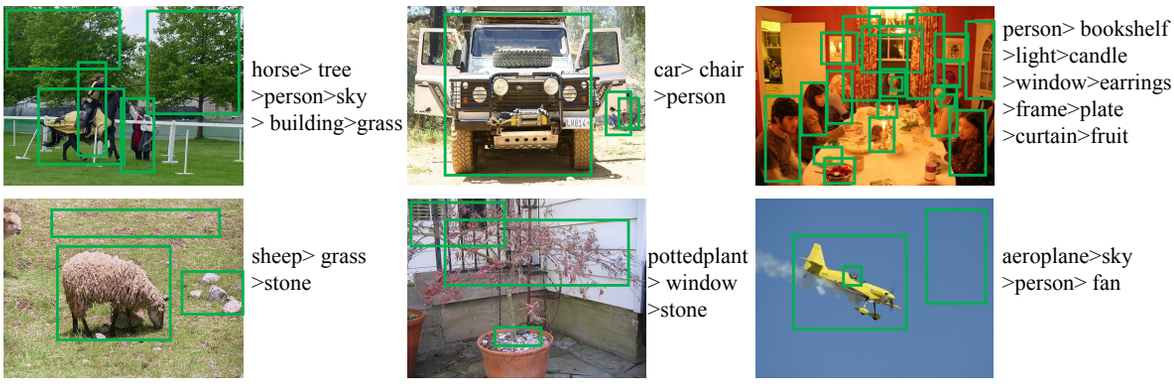


Figure 7.2: Example PASCAL images with relative importance of the objects collected on Mechanical Turk.

Thus, in this chapter, we aim to address above limitations via introducing a new method, V-SVR+, that enables the use of ranking forms of privileged information in prediction.The implicit cues of importance among the objects in the images are beneficial for multiple object recognition, as shown in Figure 7.2. We regarded the implicit cues of importance as privileged information obtained from the image absolute tag rank.

7.2 Methodology

Before proposing our method, this section first formalizes the problem of learning regression model from the labeled data along with privileged information. The goal of regression learning is to learn function $f : \mathbf{X} \rightarrow \mathbf{Y}$, which mapping from the feature space $\mathcal{X}$ to a label space $\mathcal{Y}$.

Denote a set of triples $S = \{\mathbf{X}, \mathbf{X}^*, \mathbf{Y}\}$, where $\mathbf{X} \in \mathbb{R}^{N*d}$ is a feature matrix and $\mathbf{X}^* \in \mathbb{R}^{M*d^*}$ is a privileged information matrix. N is the number of training instances, while M is the number of samples with privileged information. d and d^* are the dimensions of feature and privileged information respectively. $\mathbf{Y} \in \mathbb{R}^N$ is the ground-truth vector. When $M = N$, all privileged information can be used and we set this as the case of full privileged information. In case of existing missing privileged information, i.e. $M < N$, there is only a portion of it available for use. For null case, $M = 0$ means that no privileged information is utilized. The objective of learning with ordinal privileged information is defined as follow.

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \mathcal{L}(g(\mathbf{X_i}; \mathbf{w}), g^*(\mathbf{X_i^*}; \mathbf{w}^*), Y_i) + \alpha \sum_{i=1}^{M} \mathcal{R}(\mathbf{X_i}, \mathbf{X_i^*}, Y_i + \beta \sum_{i=1}^{M} \mathcal{V}(\mathbf{X_i^*}) \tag{7.1}$$

where $\mathcal{L}(g(\mathbf{X_i}; \mathbf{w}), g^*(\mathbf{X_i^*}; \mathbf{w}^*), Y_i)$ are the basic loss functions parameterized by $\mathbf{w}$ and $\mathbf{w}^*$, $\mathcal{R}(\mathbf{X_i}, \mathbf{X_i^*}, Y_i)$ and $\mathcal{V}(\mathbf{X_i^*})$ reflect the constraints of privileged information with ordinal (ranking) forms and numeric forms, respectively. α and β are the weighted parameters.

7.2.1 Learning with Ordinal Privileged Information

This section proposes our model to reformulate Eq. (7.1) that can address ordinal form of privileged information. More importantly, our model can be further efficiently optimized by ADMM algorithm. For the ordinal form of privileged information, we consider two ordinal forms according to different tasks, i.e. monotonically ascending ranking and monotonically descending ranking. Hence, we give the following ordinal sets of privileged information: $\mathbf{E} = \{(i, j) | X_i^* \leq X_j^* \text{ or } X_i^* > X_j^*\}$,

where $i, j \in \{1, 2..., d^*\}$. Specifically, (i, j), on one hand, when $X_i^* \leq X_j^*$, it denotes ascending relationship among privileged information. On the other hand, when $X_i^* > X_j^*$, (i, j) denotes the descending relationship.

We first use constraints between the regression classifier from general features and that from privileged information. Thus, we train two regression classifiers for these two simultaneously during training. We define two mappings from general features and privileged information to the corresponding labels as follows:

$$\begin{aligned} f(X_i) &= (w, \phi(X_i)) + b, \\ f^*(X_i^*) &= (w^*, \phi(X_i^*)) + b^*, \end{aligned} \tag{7.2}$$

where w and w^* are the weight coefficients of the general features regressor and privileged information regressor, respectively, b and b^* are their bias. $\phi(X_i)$ and $\phi(X_i^*)$ project X_i and X_i^* into the kernel space, and $(,)$ indicates the matrix product of two terms in the bracket. Then, we have the constraints as follows:

$$|Y_i - f(X_i)| \leq \epsilon + \xi_i, \tag{7.3}$$

where ξ is the slack variable that measures the failure to meet ϵ distance. Here we define $\xi_i = f^*(X_i^*)$, then the constraint of $\{f, f^*\}$ with numeric forms of privileged information is shown as Eq. (7.4)

$$\begin{aligned} f(X_i) - f^*(X_i^*) &\leq \epsilon + Y_i, \\ -f(X_i) - f^*(X_i^*) &\leq \epsilon - Y_i, \end{aligned} \tag{7.4}$$

To exploit the ordinal constraint of privileged information, we consider the defined set **E** and give the following inequation:

$$f(X_i) - f(X_j) \geq 1 - \xi_{ij}, \tag{7.5}$$

where ξ_{ij} is the slack variable that counts the number of disorder that the important privileged information is ranked below the unimportant privileged information. i and j satisfy ordinal sets **E**.

Thus, the objective function of the proposed method is as follows:

$$\begin{aligned}
&\min_{\theta,\eta,\xi} \frac{1}{2}(||\mathbf{w}||^2 + ||\mathbf{w}^*||^2) + \gamma_1 \sum_{n=1}^{N}\sum_{k=1}^{M}(\eta_{kn}^+ + \eta_{kn}^-) + \\
&\quad \gamma_2 \sum_{n=1}^{N}\sum_{(i,j)\in\mathbf{E}} \xi_{ijn} + \gamma_3 \sum_{i=1}^{M}[(\mathbf{w}^*, \mathbf{\Phi}(\mathbf{X}_\mathbf{i}^*)) + \mathbf{b}^*] \qquad (7.6)\\
&\text{s.t. } \mathbf{w}^T\phi(\mathbf{X}_{kn}) + b - Y_{kn} \le \epsilon + \eta_{kn}^+ \\
&\quad Y_{kn} - \mathbf{w}^T\phi(\mathbf{X}_{kn}) - b \le \epsilon + \eta_{kn}^- \\
&\quad \mathbf{w}^{*\mathbf{T}}\phi(\mathbf{X^*}_{kn}) + b^* - Y_{kn} \le \epsilon + \eta_{kn}^+ \\
&\quad Y_{kn} - \mathbf{w}^{*T}\phi(\mathbf{X^*}_{kn}) - b^* \le \epsilon + \eta_{kn}^- \\
&\quad \mathbf{w}^T(\phi(\mathbf{X}_{in}) - \phi(\mathbf{X}_{jn})) \ge 1 - \alpha_{ij}\xi_{ijn} \\
&\quad |\mathbf{w}^T\phi(\mathbf{X}_{kn}) + b - Y_{kn}| \le \epsilon + \mathbf{w}^{*\mathbf{T}}\phi(\mathbf{X^*}_{kn}) + b^* \\
&\quad \eta_{kn}^+, \eta_{kn}^-, \xi_{ijn}, \mathbf{w}^{*\mathbf{T}}\phi(\mathbf{X^*}_{kn}) + b^* \ge 0 \\
&\quad k = 1, ..., M, (i,j) \in \mathbf{E}, n = 1, ..., N
\end{aligned}$$

where $\theta = [\mathbf{w}; \mathbf{w}^*; b; b^*]$ is a set of parameters to be optimized. The first two constraints are used to fit regression model to the ground-truth labels. The third constraint is used to exploit ordinal relations of privileged information. The additional parameter $\alpha = \{\alpha_{ij}\} = \{\frac{1}{|i-j|}\}, \forall i \neq j$ is used to satisfy temporal smoothness.

7.2.2 Optimization of the Proposed Method

The complexity for solving in Eq. (7.6) is $O(n^2)$, n is the number of samples. In our experiment, n is usually at the level of 10^4, resulting the number of constraint 10^8. To handle optimization problem with large number of constraints, we resort to ADMM due to its compactness in solving this problem and the usage of augmented Lagrangian multipliers can also improve convergence rate. [Boyd et al., 2011].

Since ADMM can not be directly applied for Eq. (7.6), we first transform Eq. (7.6). Define $\mu = [\gamma_1; \gamma_2\alpha; \gamma_3]$, γ_1 is the coefficient of the second constraint, $\gamma_2\alpha$ is the coefficient of third constraint and γ_3 is the coefficient of fourth constraint in Eq. (7.6). $\mu \in \mathbb{R}^{M_1+M_2+2N}$ is a vector whose first M_1 entries are γ_1, the second M_2 entries are $\gamma_2\alpha$ and the last γ_3 entries are $2N$. Then, let $t = [\eta_n^+, \eta_n^-, \xi_{ijn}, \xi_{in}] \in R^{M_1+M_2+2N}$, where M_1 is the number of constraints corresponding to $[\eta_n^+, \eta_n^-]$ loss, M_2 is the number of constraints corresponding to ξ_{ijn} loss and ξ_{in} corresponds to 2N, N is the number of samples. The symbol $\lfloor . \rfloor_0$ denotes a negative value to 0. Based on the above

definitions, we can formulate Eq. (7.6) as

$$\begin{aligned} &\min_{\theta,t} \frac{1}{2}\theta^T \Lambda \theta + \mu^T \lfloor t \rfloor_0 \\ &\text{s.t. } A\theta - t = c \end{aligned} \tag{7.7}$$

where $\Lambda \in \mathbb{R}^{(d+d^*+2)\times(d+d^*+2)}$, $A \in \mathbb{R}^{(M_1+M_2+2N)\times(d+d^*+2)}$ are matrixs, and $c \in \mathbb{R}^{M_1+M_2+2N}$. Specifically,

$$A = \begin{bmatrix} \mathbf{X} & \mathbf{X}^* & \mathbf{1} & \mathbf{1} \\ -\mathbf{X} & \mathbf{X}^* & -\mathbf{1} & -\mathbf{1} \\ -\mathbf{X}_E & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{X} & -\mathbf{X}^* & \mathbf{1} & -\mathbf{1} \\ -\mathbf{X} & -\mathbf{X}^* & -\mathbf{1} & -\mathbf{1} \end{bmatrix}, c = \begin{bmatrix} \epsilon\mathbf{1} + \mathbf{y} \\ \epsilon\mathbf{1} - \mathbf{y} \\ -\mathbf{1} \\ \epsilon\mathbf{1} + \mathbf{y} \\ \epsilon\mathbf{1} - \mathbf{y} \end{bmatrix} \tag{7.8}$$

where $\mathbf{X}$ and $\mathbf{X}^*$ are the matrix of features and privileged information, respectively, and $\mathbf{X}_E$ is a matrix whose rows are the difference between two features whose indices belong to $\mathbf{E}$ set. $\mathbf{y}$ is a vector of known labels. $\mathbf{1}$ and $\mathbf{0}$ are vectors with proper dimension containing all 1s and 0s respectively.

The augmented Lagrangian has quadratic form with respect to θ, t and is linear to v. The augmented Lagrangian can be formulated as the following equation solved by ADMM.

$$\begin{aligned} L_p(\theta, t, v) =& \frac{1}{2}\theta^T \Lambda \theta + \mu^T \lfloor t \rfloor_0 + v^T (A\theta - t - c) \\ &+ \frac{\rho}{2} ||A\theta - t - c||_2^2 \end{aligned} \tag{7.9}$$

Then, the gradient of $L_p(\theta, t, v)$ w.r.t. θ or t can be computed as follows, respectively.

$$\frac{\partial L_p(\theta, t, v)}{\partial \theta} = \theta\Lambda + v^T A + \rho A^T ||A\theta - t - c||_1 \tag{7.10}$$

$$\frac{\partial L_p(\theta, t, v)}{\partial t} = \mu^T - v^T - \rho A ||A\theta - t - c||_1 \tag{7.11}$$

The new iterations of θ, t and v can be produced by letting $\partial L_p(\theta, t, v)/\partial\theta = 0$ and $\partial L_p(\theta, t, v)/\partial t = 0$.

$$\theta^{k+1} := [\frac{1}{\rho}\Lambda + A^T A]^{-1} A^T (t^k - \frac{1}{\rho} v^k + c) \tag{7.12a}$$

$$t_i^{k+1} := \frac{1}{\rho} v^k + A\theta^{k+1} - \frac{1}{\rho}\mu - c \tag{7.12b}$$

$$v^{k+1} := v^k + \rho(A\theta^{k+1} - t^{k+1} - c) \tag{7.12c}$$

In particular, due to $\lfloor t \rfloor_0 = [\eta_n^+, \eta_n^-, \xi_{ijn}, \xi_{in}]$ belonging to non-negative, we set a

threshold λ to make sure $\lfloor t \rfloor_0 \geq 0$. Let $u_i = \frac{1}{\rho}v^k + A\theta^{k+1} - \frac{1}{\rho}\mu - c$, then

$$\mathbf{t}_i^{k+1} = T_\lambda(u_i) = \begin{cases} x_i - \lambda, & \text{if} \quad u_i > \lambda \\ 0, & \text{if} \quad |u_i| \leq \lambda \\ x_i + \lambda, & \text{if} \quad u_i < -\lambda \end{cases} \tag{7.13}$$

where $\lambda \geq 0$ is the soft threshold operator and i is the ith entry in each vector. The pseudo-code of our method is shown in Algorithm 4. The algorithm will terminate if the error becomes smaller than some tolerance value or the maximum number of iterations reaches. After solving the optimization problem, we can predict unknown labels for new samples x as $y = \mathbf{w}^T x + b$.

Algorithm 4 RaPI learning by ADMM

Require: Feature information $\mathbf{X}$, Privileged information $\mathbf{X}^*$, Labels $\mathbf{Y}$, pairwise index set $\mathbf{E}$
Ensure: Model parameters $\{\theta, t, v\}$
Construct Λ, μ, A, c using $\mathbf{X}, \mathbf{X}^*, \mathbf{Y}, \mathbf{E}$
$\theta \leftarrow 0, t \leftarrow 0, v \leftarrow 0$
repeat
Update θ^{k+1}, t^{k+1} and v^{k+1} by Eq. (7.12)
until
$t^{k+1} - t^k \rightarrow 0$
$A\theta^{k+1} - t^{k+1} - c \rightarrow 0$
return θ

7.3 Experiments

7.3.1 Experimental Conditions

7.3.1.1 Dataset

The Pascal VOC 2007 database contains 9963 images with 20 classes as target labels. The database is divided into two subsets: 5011 samples for training and 4952 samples for testing. We use 512-dimensional gist features, 200-dimensional bag of visual word features [Oliva and Torralba, 2006][Hwang and Grauman, 2012] and 64 dimensional color histogram features [Hwang and Grauman, 2010][Hwang and Grauman, 2012] in the experiments. Specifically, the 339-dimensional absolute tag rank features used as privileged information during training are provided in [Hwang and Grauman, 2010] where 20 classes as target labels are given with continuous values. Thus, we can use a regression method to finish this classification task.

7.3.1.2 Experimental Design

For multiple object recognition from images with the help of implicit information, we conduct experiments on 20 category labels. Although multiple object recognition is a classification task, we use the regression method to complete this task because 20 category labels have been given [Hwang and Grauman, 2012], and the deviation degree of each predicted classification and the ground truth must be calculated. Thus, we conduct the following experiments: using features without any implicit importance cues (**SVM**) and using features with the help of implicit importance cues as privileged information (**V-SVR+**).

7.3.1.3 Performance Metrics

We use the average precision (AP) as the metric used in the PASCAL challenge [1][Everingham et al., 2010].

7.3.2 Experimental Results and Analysis

Table 7.1: Average precision (AP) scores per-class results on the Pascal VOC 2007 database. The results of the proposed method rank first in 11 out of 20 classes.

Methods	bicycle	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motorbike	person	potted plant	sheep	sofa	train	tvmonitor
latent SVM	38.83	38.86	65.77	12.14	19.10	36.91	55.62	55.10	29.33	22.31	34.31	45.80	14.81	61.80	52.90	38.95	18.81	32.55	40.17	49.87	51.58
tags_LSVM	39.51	38.99	65.87	12.56	20.16	30.07	57.17	55.26	32.41	22.49	35.17	45.04	16.64	61.31	53.38	38.61	21.77	32.94	40.52	50.88	51.96
Context-DA	64	68.9	73.1	**62.5**	57.6	38.9	**72.5**	74.8	**77.2**	42.9	**69.7**	59.5	**63.9**	76.1	70.2	69.2	**43.9**	**58.3**	59.7	77.2	**64.8**
winner VOC 2007	59.4	77.5	63.6	56.1	71.9	33.1	60.6	78.0	55.8	53.5	42.6	54.9	45.8	77.5	64.0	**85.9**	36.3	44.7	50.9	79.2	53.2
SVR	55.7	74.2	70.3	52.7	66.4	38.2	46.3	68.5	51.2	43.1	42.6	53.2	50.7	71.6	59.2	64.9	32.8	52.5	49.8	70.6	55.4
V-SVR+	**65.2**	**85.2**	**74.8**	59.8	**72.4**	**39.8**	68.8	**81.3**	62.8	**58.7**	52.9	**64.2**	55.8	**79.4**	**70.8**	77.6	35.7	57.2	**62.4**	**82.3**	62.8
Our rank	1	1	1	2	1	1	2	1	2	1	2	1	2	1	1	2	3	2	1	1	2

With the Pascal VOC 2007 database, the recognition performance of every label is shown in Table 7.1 where we list our scores for all 20 classes. We compare the proposed approach (V-SVR+) with the approach without privileged information (SVR), the other three methods and the best performance of the challenge (winner VOC 2007). As shown in Table 7.1, the proposed method achieves better than the SVR method. Our method also achieves better results than those achieved in the official competition for 18 of 20 categories. For many objects (aeroplane, bicycle, bottle, bus, cat, cow, dining table, dog, motorbike, sheep, sofa and TVmonitor), we see good improvements. The proposed method also achieves better performance on mean AP, which indicates that ordinal privileged information contribute to better multi-label classification.

[1]http://host.robots.ox.ac.uk/pascal/VOC/voc2007/

7.3.3 Comparison to Related Works

We use the Pascal VOC 2007 database to demonstrate the superiority of the proposed regression method with ordinal privileged information. Related studies are described below:

- Latent SVM [Felzenszwalb et al., 2008] combines a margin-sensitive approach for data mining hard negative examples with a formalism called latent SVM, to allow the effective use of more latent information for object detection tasks.
- Tags_LSVM [Hwang and Grauman, 2011] combines SVM and three novel implicit features from an image tag to improve both accuracy and efficiency when detecting the tagged objects.
- The Context-DA method [Dvornik et al., 2019] uses a convolutional neural network to improve object detection and a CNN network is used to predict the suitability of an image region for placing a given object.

A comparison of results is shown in Table 7.1. Compared to the other three methods, the proposed method obtains the highest score in 12 categories, the second-highest score in 8 categories and the best performance on mean AP. We compare the proposed method with the latent SVM and tags_LSVM because the proposed method is based on a support vector regression. We also compare the proposed method with the content-DA [Dvornik et al., 2019] model in the single-category experiment. Both the Content-DA model and the proposed are trained independently for each category rather than all categories at the same time. The proposed method can achieve more improvements by considering the dependencies among multiple labels. Because this is not the focus of this paper, we will address this topic in future work.

7.4 Conclusion

In this paper, we propose a novel V-SVR+ model with ordinal constraint to capture the effect of potential information for prediction. V-SVR+ model can address ordinal form of privileged information by integrating ordinal constraints into the learning process. Experimental results multiple object recognition show that the implicit information about object importance considered to be ordinal privileged information is beneficial for finishing a better prediction task and V-SVR+ model involving ordinal form of privileged information can benefit the regression learning.

Chapter 8

Conclusion

8.1 Summary

In this thesis, we exploit the asymmetric distribution of some additional information only available in training in machine learning for prediction. We introduce how to learn under variational privileged information by leveraging asymmetric distribution, and combine some machine learning algorithms such as semi-supervised learning, adversarial learning, multi-task learning and support vector regression. We also use three different tasks, music emotion recognition, multiple object recognition, and photo aesthetic assessment to demonstrate the proposed methods.

1. Photo aesthetic assessment: we propose a deep multi-task network that utilizes high-level descriptive aesthetic attributes as privileged information to enhance the photo assessment. Pairwise photos are first trained respectively via a basic generator network. During the fine-tuning process of generative network, the ranking constraint is introduced in the label space to capture the ordinal relation between attributes and aesthetic score. Adversarial learning is used to enforce similarity between join distribution of the predicted attributes and ground truth attributes; prior probabilities loss is used to exploit the coexistent and exclusive relations among attributes; and similarity constraint are introduced to further regularize the learned aesthetic attributes. Our method achieves state-of-the-art performance on the task of photo aesthetic assessment on two benchmark databases, demonstrating that the proposed method captures the correlation between aesthetics and variational privileged information hidden in photos, and then can benefit the aesthetic assessment.

2. Music emotion recognition: we summarize the domain knowledge of musicology and exploit the relations among musical elements, music styles and musical emotion. We first propose the use of RBM model to capture their relations and construct a better privileged information representation. Then we use the generated new privileged information to estimate the slack variate of support vector regression involving the learning process and propose the novel SVR+ methods with musical

elements and music styles of two kinds of musicology knowledge for music emotion recognition. Finally, we provide a detailed optimization algorithm to solve the proposed model. We conduct extensive experiments to demonstrate the superiority of the proposed SVR+ with privileged information method on the MediaEval 2015 and AMG 1608 database.

3. Multiple object recognition: we propose a novel model with ranking constraint to capture the effect of potential information. The proposed model can address ordinal form of privileged information by integrating ranking constraints into the learning process. Experimental results on multiple object recognition show that the implicit information about object importance is considered to be ordinal privileged information, which is beneficial for finishing a better recognition task.

8.2 Future work

This thesis proposes some methods of privileged machine learning for prediction, and make some progress on photo aesthetic assessment, music emotion recognition and multiple object recognition. However, we consider further research in the following aspects:

1. In the future, we will exploit privileged information, including posterior information, future information and hidden information etc. for choosing the best hyper-parameters of the neural network model to further optimize the performance of our approach. With the development of deep learning, choosing the suitable hyper-parameters is a increasing demand. Privileged information is able to provide some guidance and explanation for hyper-parameters selection, which will increase the efficiency of deep network.

2. In this thesis, we find high-level descriptive attributes are importance and can be used as privileged information on photo aesthetic assessment. However, the attribute of "content" have an impact on human's aesthetic judgement due to heavily relying on human preference. How to eliminate visual preference becomes an important topic and will be exploited in the future work.

3. This thesis aims to music emotion recognition with some musical pieces. In the future work, It would be better to combine the current audio feature MER with video affective content analysis, such as music in movies and music in videoland. The proposed method also can be extended to many fields via SVR+ or SVM+ model with domain knowledge as privileged information such as video sentiment analysis. We will explore these in future work.

4. This thesis proposes to use the implicit information about object importance as ordinal privileged information for multiple object recognition. However, we ignore

the relations among multiple objects. In the future, we exploit the label correlations or object correlations for better object recognition.

Bibliography

Aljanaki, A.; Yang, Y.-H.; and Soleymani, M., 2015. Emotion in music task at mediaeval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop*. (cited on pages 65, 67, and 69)

Bahuleyan, H., 2018. Music genre classification using machine learning techniques. *arXiv preprint arXiv:1804.01149*, (2018). (cited on page 55)

Barrett, L. F., 1998. Discrete emotions or dimensions? the role of valence focus and arousal focus. *Cognition & Emotion*, 12, 4 (1998), 579–599. (cited on page 51)

Best, J. A.; Grimsley, J. F.; Heise, A. C.; Mascioli, L. M.; Moroney, S. F.; Satterfield, L. A.; and Zeller, E. I., 2014. Global music perspectives: Music outside the western canon in local schools. (2014). (cited on page 56)

Bohlman, P. V., 1988. *The study of folk music in the modern world*. Indiana University Press. (cited on page 56)

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3, 1 (2011), 1–122. (cited on pages 3, 5, 17, 22, 23, and 79)

Cabannes, V.; Bach, F.; and Rudi, A., 2021. Disambiguation of weak supervision with exponential convergence rates. *arXiv preprint arXiv:2102.02789*, (2021). (cited on page 15)

Cai, F., 2011. Advanced learning approaches based on svm+ methodology. (2011). (cited on page 8)

Caruana, R., 1997. Multitask learning. *Machine learning*, 28, 1 (1997), 41–75. (cited on pages 4 and 16)

Chapaneri, S. and Jayaswal, D., 2017. Structured prediction of music mood with twin gaussian processes. In *International Conference on Pattern Recognition and Machine Intelligence*, 647–654. Springer. (cited on page 69)

Chen, C.; Wu, Z.; and Jiang, Y.-G., 2016. Emotion in context: Deep semantic feature fusion for video emotion recognition. In *Proceedings of the 24th ACM international conference on Multimedia*, 127–131. (cited on page 11)

Chen, Y.-A.; Wang, J.-C.; Yang, Y.-H.; and Chen, H., 2014. Linear regression-based adaptation of music emotion recognition models for personalization. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2149–2153. IEEE. (cited on page 52)

Chen, Y.-A.; Yang, Y.-H.; Wang, J.-C.; and Chen, H., 2015. The amg1608 dataset for music emotion recognition. In *ICASSP 2015*, 693–697. IEEE. (cited on pages 65 and 66)

Chmulik, M.; Guoth, I.; Malik, M.; and Jarina, R., 2015. Uniza system for the" emotion in music" task at mediaeval 2015. In *MediaEval*. (cited on pages 51, 68, and 69)

Choi, K.; Fazekas, G.; and Sandler, M., 2016. Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298*, (2016). (cited on page 11)

Cook, T.; Roy, A. R.; and Welker, K. M., 2019. Music as an emotion regulation strategy: An examination of genres of music and their roles in emotion regulation. *Psychology of Music*, 47, 1 (2019), 144–154. (cited on page 57)

Coutinho, E.; Trigeorgis, G.; Zafeiriou, S.; and Schuller, B., 2015. Automatically estimating emotion in music with deep long-short term memory recurrent neural networks. In *CEUR Workshop Proceedings*, vol. 1436. (cited on pages 68 and 69)

Cui, C.; Liu, H.; Lian, T.; Nie, L.; Zhu, L.; and Yin, Y., 2018. Distribution-oriented aesthetics assessment with semantic-aware hybrid network. *IEEE Transactions on Multimedia*, 21, 5 (2018), 1209–1220. (cited on pages 9, 27, and 28)

Deng, J. J.; Leung, C. H.; Milani, A.; and Chen, L., 2015. Emotional states associated with music: Classification, prediction of changes, and consideration in recommendation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5, 1 (2015), 1–36. (cited on page 10)

Deng, Y.; Loy, C. C.; and Tang, X., 2017. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34, 4 (2017), 80–106. (cited on page 8)

Dhar, S.; Ordonez, V.; and Berg, T. L., 2011. High level describable attributes for predicting aesthetics and interestingness. In *CVPR 2011*, 1657–1664. IEEE. (cited on pages 31 and 75)

Dong, J. and Lin, T., 2019. Margingan: Adversarial training in semi-supervised learning. *Advances in neural information processing systems*, 32 (2019), 10440–10449. (cited on page 15)

Drucker, H.; Burges, C. J.; Kaufman, L.; Smola, A. J.; and Vapnik, V., 1997. Support vector regression machines. In *Advances in neural information processing systems*, 155–161. (cited on page 3)

Dvornik, N.; Mairal, J.; and Schmid, C., 2019. On the importance of visual context for data augmentation in scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2019). (cited on pages 12 and 83)

Eerola, T., 2011. Are the emotions expressed in music genre-specific? an audio-based evaluation of datasets spanning classical, film, pop and mixed genres. *Journal of New Music Research*, 40, 4 (2011), 349–366. (cited on page 12)

Erlewine, M., 1999. *All music guide to the blues: the experts' guide to the best blues recordings*. Backbeat Books. (cited on page 56)

Everett III, H., 1963. Generalized lagrange multiplier method for solving problems of optimum allocation of resources. *Operations research*, 11, 3 (1963), 399–417. (cited on pages 3, 17, 19, and 61)

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A., 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88, 2 (2010), 303–338. (cited on page 82)

Eyben, F.; Wöllmer, M.; and Schuller, B., 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, 1459–1462. (cited on pages 50 and 65)

Felzenszwalb, P.; McAllester, D.; and Ramanan, D., 2008. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. IEEE. (cited on pages 12 and 83)

Fernández-Sotos, A.; Fernández-Caballero, A.; and Latorre, J. M., 2016. Influence of tempo and rhythmic unit in musical emotion regulation. *Frontiers in computational neuroscience*, 10 (2016), 80. (cited on page 53)

Friberg, A.; Schoonderwaldt, E.; Hedblad, A.; Fabiani, M.; and Elowsson, A., 2014. Using listener-based perceptual features as intermediate representations in music information retrieval. *The Journal of the Acoustical Society of America*, 136, 4 (2014), 1951–1963. (cited on pages 50 and 51)

Frith, S. et al., 2001. *The Cambridge companion to pop and rock*. Cambridge University Press. (cited on page 56)

Gabrielsson, A. and Lindström, E., 2010. The role of structure in the musical expression of emotions. *Handbook of music and emotion: Theory, research, applications*, 367400 (2010). (cited on pages 53, 54, and 55)

Gomez, P. and Danuser, B., 2007. Relationships between musical structure and psychophysiological measures of emotion. *Emotion*, 7, 2 (2007), 377. (cited on page 53)

Goodfellow, I., 2016. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, (2016). (cited on page 34)

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y., 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680. (cited on pages 3, 14, 15, 33, 34, and 38)

Han, Y.; Kim, J.; and Lee, K., 2016. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25, 1 (2016), 208–221. (cited on page 11)

He, D.; Xia, Y.; Qin, T.; Wang, L.; Yu, N.; Liu, T.-Y.; and Ma, W.-Y., 2016a. Dual learning for machine translation. *Advances in neural information processing systems*, 29 (2016), 820–828. (cited on page 6)

He, K.; Zhang, X.; Ren, S.; and Sun, J., 2016b. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778. (cited on page 40)

Hevner, K., 1937. The affective value of pitch and tempo in music. *The American Journal of Psychology*, 49, 4 (1937), 621–630. (cited on page 55)

Heyman, B. B., 2019. *Samuel Barber: the composer and his music*. Oxford University Press. (cited on page 56)

Hinton, G. E., 2012. A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade*, 599–619. Springer. (cited on page 60)

Hinton, G. E. and Salakhutdinov, R. R., 2006. Reducing the dimensionality of data with neural networks. *science*, 313, 5786 (2006), 504–507. (cited on page 59)

Holmes, T. B. and Holmes, T., 2002. *Electronic and experimental music: pioneers in technology and composition*. Psychology Press. (cited on page 56)

Hosu, V.; Goldlucke, B.; and Saupe, D., 2019. Effective aesthetics prediction with multi-level spatially pooled features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9375–9383. (cited on pages 27, 41, and 42)

Hou, L.; Yu, C.-P.; and Samaras, D., 2016. Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*, (2016). (cited on page 42)

Hu, X. and Yang, Y.-H., 2017. Cross-dataset and cross-cultural music mood prediction: A case on western and chinese pop songs. *IEEE Transactions on Affective Computing*, 8, 2 (2017), 228–240. (cited on pages 11, 52, 68, and 69)

Husain, G.; Thompson, W. F.; and Schellenberg, E. G., 2002. Effects of musical tempo and mode on arousal, mood, and spatial abilities. *Music perception*, 20, 2 (2002), 151–171. (cited on page 53)

Hwang, S. J. and Grauman, K., 2010. Accounting for the relative importance of objects in image retrieval. In *BMVC*, vol. 1, 5. (cited on page 81)

Hwang, S. J. and Grauman, K., 2011. Reading between the lines: Object localization using implicit cues from image tags. *IEEE transactions on pattern analysis and machine intelligence*, 34, 6 (2011), 1145–1158. (cited on pages 12 and 83)

Hwang, S. J. and Grauman, K., 2012. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *International journal of computer vision*, 100, 2 (2012), 134–153. (cited on pages 81 and 82)

Ilie, G. and Thompson, W. F., 2006. A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception*, 23, 4 (2006), 319–330. (cited on page 55)

Jiang, Y., 2004. *The Chinese violin concerto "The Butterfly Lovers" by He Zhanhao (1933) and Chen Gang (1935) for violin and orchestra*. Ph.D. thesis, The University of Texas at Austin. (cited on page 51)

JLB, D. P. K., 2015. Adam: A method for stochastic optimization. In *3rd international conference for learning representations, San Diego*. (cited on page 40)

Jones, G. T., 1974. *Music Theory: The Fundamental Concepts of Tonal Music Including Notation, Terminology, and Harmony*. (cited on page 54)

Joshi, D.; Datta, R.; Fedorovskaya, E.; Luong, Q.-T.; Wang, J. Z.; Li, J.; and Luo, J., 2011. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28, 5 (2011), 94–115. (cited on page 8)

Kamienny, P.-A.; Arulkumaran, K.; Behbahani, F.; Boehmer, W.; and Whiteson, S., 2020. Privileged information dropout in reinforcement learning. *arXiv preprint arXiv:2005.09220*, (2020). (cited on page 13)

Kao, Y.; He, R.; and Huang, K., 2017. Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing*, 26, 3 (2017), 1482–1495. (cited on pages 9, 28, and 42)

Ke, Y.; Tang, X.; and Jing, F., 2006. The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, 419–426. IEEE. (cited on page 9)

Kong, S.; Shen, X.; Lin, Z.; Mech, R.; and Fowlkes, C., 2016. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, 662–679. Springer. (cited on pages 10, 27, 39, and 41)

Koutras, A., 2017. Song emotion recognition using music genre information. In *International Conference on Speech and Computer*, 669–679. Springer. (cited on page 12)

Lai, W.-S.; Huang, J.-B.; and Yang, M.-H., 2017. Semi-supervised learning for optical flow with generative adversarial networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 353–363. (cited on pages 4 and 15)

Lambert, J.; Sener, O.; and Savarese, S., 2018. Deep learning under privileged information using heteroscedastic dropout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8886–8895. (cited on pages 2, 13, and 52)

Lapin, M.; Hein, M.; and Schiele, B., 2014. Learning using privileged information: Svm+ and weighted svm. *Neural Networks*, 53 (2014), 95–108. (cited on pages 1 and 7)

Lartillot, O., 2011. Mirtoolbox 1.3. 4 user's manual. *Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Finland*, (2011). (cited on pages 51, 53, and 66)

Li, L.; Zhu, H.; Zhao, S.; Ding, G.; Jiang, H.; and Tan, A., 2019. Personality driven multi-task learning for image aesthetic assessment. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 430–435. IEEE. (cited on page 41)

Li, W.; Dai, D.; Tan, M.; Xu, D.; and Van Gool, L., 2016a. Fast algorithms for linear and kernel svm+. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2258–2266. (cited on pages 1 and 8)

Li, W.; Niu, L.; and Xu, D., 2014. Exploiting privileged information from web data for image categorization. In *European Conference on Computer Vision*, 437–452. Springer. (cited on page 13)

Li, X.; Du, B.; Xu, C.; Zhang, Y.; Zhang, L.; and Tao, D., 2018. R-svm+: Robust learning with privileged information. In *IJCAI*, 2411–2417. (cited on page 8)

Li, X.; Xianyu, H.; Tian, J.; Chen, W.; Meng, F.; Xu, M.; and Cai, L., 2016b. A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 544–548. IEEE. (cited on pages 68 and 69)

Lightstone, A., 2012. The importance of hip-hop for music therapists. *Therapeutic uses of rap and hip-hop*, (2012), 39–56. (cited on page 56)

Liu, D.; Puri, R.; Kamath, N.; and Bhattacharya, S., 2020. Composition-aware image aesthetics assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3569–3578. (cited on page 42)

Liu, Y.; Liu, Y.; and Gu, Z., 2015. Affective feature extraction for music emotion prediction. In *MediaEval*. (cited on page 51)

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60, 2 (2004), 91–110. (cited on page 8)

Lu, X.; Lin, Z.; Jin, H.; Yang, J.; and Wang, J. Z., 2014. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, 457–466. (cited on page 9)

Luo, Y. and Tang, X., 2008. Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision*, 386–399. Springer. (cited on pages 9 and 27)

Lv, P.; Wang, M.; Xu, Y.; Peng, Z.; Sun, J.; Su, S.; Zhou, B.; and Xu, M., 2018. Usar: an interactive user-specific aesthetic ranking framework for images. *arXiv preprint arXiv:1805.01091*, (2018). (cited on pages 27 and 28)

Malik, M.; Adavanne, S.; Drossos, K.; Virtanen, T.; Ticha, D.; and Jarina, R., 2017. Stacked convolutional and recurrent neural networks for music emotion recognition. *arXiv preprint arXiv:1706.02292*, (2017). (cited on page 51)

Malu, G.; Bapi, R. S.; and Indurkhya, B., 2017. Learning photography aesthetics with deep cnns. *arXiv preprint arXiv:1707.03981*, (2017). (cited on pages 9, 28, and 41)

Mao, Q.; Dong, M.; Huang, Z.; and Zhan, Y., 2014. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE transactions on multimedia*, 16, 8 (2014), 2203–2213. (cited on page 11)

Markov, K. and Matsui, T., 2014. Music genre and emotion recognition using gaussian processes. *IEEE access*, 2 (2014), 688–697. (cited on page 52)

Miller, M., 2005. *The complete idiot's guide to music theory*. Penguin. (cited on page 54)

Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S., 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41, 8 (2018), 1979–1993. (cited on page 15)

Motiian, S., 2019. Domain adaptation and privileged information for visual recognition. (2019). (cited on page 13)

Murray, N.; Marchesotti, L.; and Perronnin, F., 2012. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2408–2415. IEEE. (cited on page 39)

Nishiyama, M.; Okabe, T.; Sato, I.; and Sato, Y., 2011. Aesthetic quality classification of photographs based on color harmony. In *CVPR 2011*, 33–40. IEEE. (cited on page 28)

Niu, L. and Wu, J., 2012. Nonlinear l-1 support vector machines for learning using privileged information. In *2012 IEEE 12th International Conference on Data Mining Workshops*, 495–499. IEEE. (cited on pages 1 and 7)

Oliva, A. and Torralba, A., 2006. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155 (2006), 23–36. (cited on page 81)

Oramas, S.; Barbieri, F.; Nieto, O.; and Serra, X., 2018. Multimodal deep learning for music genre classification. *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.*, (2018). (cited on page 55)

Orjesek, R.; Jarina, R.; Chmulik, M.; and Kuba, M., 2019. Dnn based music emotion recognition from raw audio signal. In *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, 1–4. IEEE. (cited on pages 68 and 69)

Pan, B.; Wang, S.; and Jiang, Q., 2019a. Image aesthetic assessment assisted by attributes through adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 679–686. (cited on pages 10, 28, and 41)

Pan, B.; Wang, S.; and Xia, B., 2019b. Occluded facial expression recognition enhanced through privileged information. In *Proceedings of the 27th ACM International Conference on Multimedia*, 566–573. (cited on page 13)

Panda, R.; Malheiro, R. M.; and Paiva, R. P., 2018. Novel audio features for music emotion recognition. *IEEE transactions on affective computing*, (2018). (cited on pages 11 and 50)

Panda, R.; Malheiro, R. M.; and Paiva, R. P., 2020. Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing*, (2020). (cited on pages 10 and 51)

Patra, B. G.; Das, D.; and Bandyopadhyay, S., 2013. Unsupervised approach to hindi music mood classification. In *Mining Intelligence and Knowledge Exploration*, 62–69. Springer. (cited on page 11)

Pellegrini, T. and Barriere, V., 2015. Time-continuous estimation of emotion in music with recurrent neural networks. In *MediaEval 2015 Multimedia Benchmark Workshop (MediaEval 2015)*, pp–1. (cited on pages 68 and 69)

Peterson, R. A., 1978. The production of cultural change: The case of contemporary country music. *Social Research*, (1978), 292–314. (cited on page 56)

Plack, C. J.; Oxenham, A. J.; and Fay, R. R., 2006. *Pitch: neural coding and perception*, vol. 24. Springer Science & Business Media. (cited on page 55)

Platt, J., 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. (1998). (cited on page 20)

Qi, G.-J.; Aggarwal, C. C.; and Huang, T., 2013a. Link prediction across networks by biased cross-network sampling. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 793–804. IEEE. (cited on page 13)

Qi, G.-J.; Aggarwal, C. C.; and Huang, T. S., 2013b. Online community detection in social sensing. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 617–626. (cited on page 13)

Radovic, M.; Adarkwa, O.; and Wang, Q., 2017. Object recognition in aerial images using convolutional neural networks. *Journal of Imaging*, 3, 2 (2017), 21. (cited on page 13)

Rentfrow, P. J. and Gosling, S. D., 2003. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology*, 84, 6 (2003), 1236. (cited on page 57)

Saitis, C. and Siedenburg, K., 2020. Brightness perception for musical instrument sounds: Relation to timbre dissimilarity and source-cause categories. *The Journal of the Acoustical Society of America*, 148, 4 (2020), 2256–2266. (cited on pages 12 and 51)

Sarafianos, N.; Nikou, C.; and Kakadiaris, I. A., 2016. Predicting privileged information for height estimation. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 3115–3120. IEEE. (cited on page 8)

Sarafianos, N.; Vrigkas, M.; and Kakadiaris, I. A., 2017. Adaptive svm+: Learning with privileged information for domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2637–2644. (cited on page 13)

Scherer, K. R. and Oshinsky, J. S., 1977. Cue utilization in emotion attribution from auditory stimuli. *Motivation and emotion*, 1, 4 (1977), 331–346. (cited on page 55)

Schwartz, E.; Karlinsky, L.; Shtok, J.; Harary, S.; Marder, M.; Feris, R.; Kumar, A.; Giryes, R.; and Bronstein, A. M., 2018. Delta-encoder: an effective sample synthesis method for few-shot object recognition. *arXiv preprint arXiv:1806.04734*, (2018). (cited on page 12)

Sharmanska, V.; Quadrianto, N.; and Lampert, C. H., 2013. Learning to rank using privileged information. In *Proceedings of the IEEE International Conference on Computer Vision*, 825–832. (cited on pages 2, 52, and 76)

Shu, Y.; Li, Q.; Liu, S.; and Xu, G., 2020. Learning with privileged information for photo aesthetic assessment. *Neurocomputing*, (2020). (cited on pages 8, 28, and 41)

Shu, Y.; Li, Q.; Xu, C.; Liu, S.; and Xu, G., 2021. V-svr+: Support vector regression with variational privileged information. *IEEE Transactions on Multimedia*, (2021). (cited on pages 10 and 41)

Shu, Y. and Xu, G., 2019. Emotion recognition from music enhanced by domain knowledge. In *Pacific Rim International Conference on Artificial Intelligence*, 121–134. Springer. (cited on pages 12, 68, and 69)

Starr, L. and Waterman, C., 2008. Popular jazz and swing: America's original art form. (cited on page 56)

Studwell, W. E. and Lonergan, D. F., 1999. *The classic rock and roll reader: rock music from its beginnings to the mid-1970s*. Psychology Press. (cited on page 56)

Susino, M. and Schubert, E., 2019. Cultural stereotyping of emotional responses to music genre. *Psychology of Music*, 47, 3 (2019), 342–357. (cited on pages 56 and 57)

Talebi, H. and Milanfar, P., 2018. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27, 8 (2018), 3998–4011. (cited on page 42)

Tang, J.; Hua, X.-S.; Mei, T.; Qi, G.-J.; and Wu, X., 2007a. Video annotation based on temporally consistent gaussian random field. *Electronics Letters*, 43, 8 (2007), 448–449. (cited on page 13)

Tang, J.; Hua, X.-S.; Qi, G.-J.; and Wu, X., 2007b. Typicality ranking via semi-supervised multiple-instance learning. In *Proceedings of the 15th ACM international conference on Multimedia*, 297–300. (cited on page 13)

Vapnik, V. and Vashist, A., 2009. A new learning paradigm: Learning using privileged information. *Neural networks*, 22, 5-6 (2009), 544–557. (cited on pages 1, 7, 43, 52, 57, and 61)

Wang, J.; Zhao, Z.; Zhou, J.; Wang, H.; Cui, B.; and Qi, G., 2012a. Recommending flickr groups with social topic model. *Information retrieval*, 15, 3-4 (2012), 278–295. (cited on page 13)

Wang, J.-C.; Yang, Y.-H.; and Wang, H.-M., 2016. Affective music information retrieval. In *Emotions and Personality in Personalized Services*, 227–261. Springer. (cited on pages 68 and 69)

Wang, J.-C.; Yang, Y.-H.; Wang, H.-M.; and Jeng, S.-K., 2012b. The acoustic emotion gaussians model for emotion-based music annotation and retrieval. In *Proceedings of the 20th ACM international conference on Multimedia*, 89–98. (cited on page 10)

Wang, Z.; Liu, D.; Chang, S.; Dolcos, F.; Beck, D.; and Huang, T., 2017. Image aesthetics assessment using deep chatterjee's machine. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 941–948. IEEE. (cited on page 42)

Weber, M.; Krismayer, T.; Wöss, J.; Aigmüller, L.; and Birnzain, P., 2015. Mediaeval 2015: Jku-tinnitus approach to emotion in music task. In *MediaEval*. (cited on page 51)

Wessel, D. L., 1979. Timbre space as a musical control structure. *Computer music journal*, (1979), 45–52. (cited on page 54)

Xianyu, H.; Li, X.; Chen, W.; Meng, F.; Tian, J.; Xu, M.; and Cai, L., 2016. Svr based double-scale regression for dynamic emotion prediction in music. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 549–553. IEEE. (cited on pages 11, 68, and 69)

Xu, X.; Li, W.; and Xu, D., 2015. Distance metric learning using privileged information for face verification and person re-identification. *IEEE transactions on neural networks and learning systems*, 26, 12 (2015), 3150–3162. (cited on page 2)

Yan, Y.; Nie, F.; Li, W.; Gao, C.; Yang, Y.; and Xu, D., 2016. Image classification by cross-media active learning with privileged information. *IEEE Transactions on Multimedia*, 18, 12 (2016), 2494–2502. (cited on page 13)

Yang, H.; Shi, P.; He, S.; Pan, D.; Ying, Z.; and Lei, L., 2019. A comprehensive survey on image aesthetic quality assessment. In *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, 294–299. IEEE Computer Society. (cited on page 8)

Yang, X.; Dong, Y.; and Li, J., 2018. Review of data features-based music emotion recognition methods. *Multimedia systems*, 24, 4 (2018), 365–389. (cited on page 10)

You, S.; Xu, C.; Wang, Y.; Xu, C.; and Tao, D., 2017. Privileged multi-label learning. *arXiv preprint arXiv:1701.07194*, (2017). (cited on page 2)

Zbikowski, L. M. and Zbikowski, L. M., 2002. *Conceptualizing music: Cognitive structure, theory, and analysis*. Oxford University Press on Demand. (cited on pages 51 and 53)

Zeng, H.; Cao, Z.; Zhang, L.; and Bovik, A. C., 2019. A unified probabilistic formulation of image aesthetic assessment. *IEEE Transactions on Image Processing*, 29 (2019), 1548–1561. (cited on page 42)

Zhang, X.; Gao, X.; Lu, W.; and He, L., 2019. A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction. *IEEE Transactions on Multimedia*, (2019). (cited on page 42)

Zhao, Z.-Q.; Zheng, P.; Xu, S.-t.; and Wu, X., 2019. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30, 11 (2019), 3212–3232. (cited on page 12)

Zhu, H.; Shan, H.; Zhang, Y.; Che, L.; Xu, X.; Zhang, J.; Shi, J.; and Wang, F.-Y., 2021. Convolutional ordinal regression forest for image ordinal estimation. *IEEE Transactions on Neural Networks and Learning Systems*, (2021). (cited on page 42)