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

Yangyang Shu, Qian Li, Chang Xu, Shaowu Liu, and Guandong Xu

***Abstract*—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. Therefore, in this paper, we 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. We evaluate the proposed method in three applications: music emotion recognition from songs with the help of implicit information about music elements judged by composers; 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; and photo aesthetic assessment enhanced by high-level aesthetic attributes hidden in photos. Experiment results demonstrate that the proposed methods are superior to the classic learning paradigm when solving practical problems.**

***Index Terms*—support vector regression, variational privileged information**

I. Introduction

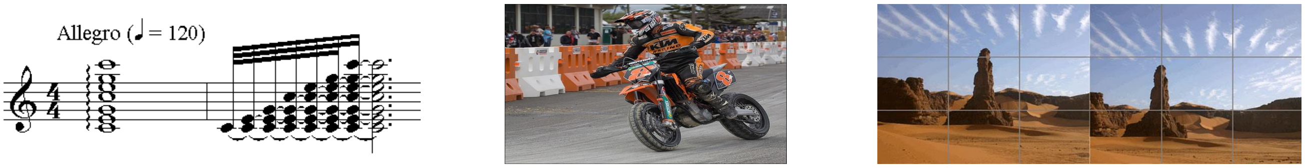
In most traditional regression algorithms, the same information distribution is required to both train and test the model. However, one can access not only the input/output training pairs of the task we want to learn but also additional information on the training examples (the so-called privileged information (PI)), which can include photo aesthetic attributes in photo assessment tasks. Typically, this additional information is more informative to learning tasks than using training data alone; thus, PI has been extensively researched to enhance various learning tasks. However, privileged information is typically not easily obtained for reuse in testing due to the higher data collection costs and/or a lack of access to domain expertise.

To address this issue, Vapnik and Vashist et al. [1] first proposed the *learning using privileged information (LUPI)* paradigm to integrate privileged information into the learning process and achieved success with various machine learning algorithms, including regression tasks, which is the focus of this paper. By exploiting privileged information, a regression model can be trained better when PI is involved during learning or parameter optimization. Taking music as an example, the tempo, mode, brightness, and loudness of music elements, which implicitly exist in songs, can improve the musical quality of each song. Because these music elements can be extracted during training, they are only available during training but not during testing. Learning from privileged information is also common in many other applications. For example, in multiple object recognition, the ordinal relations among different objects, which are implicitly supplied by annotators, imply the importance of the image's context. For photo aesthetic assessment, high-level aesthetic attributes such as the rule of third, complementary and motion blur are provided by database promulgators. Photos with these implicit high-level attributes are beneficial for aesthetics. To use the LUPI paradigm, Vapnik and Vashist [1] first upgraded the support vector machine (SVM), proposed the SVM+ method and extended the SVM+ method (e.g., a mixture model of slacks of learning using privileged information). Because SVM+ is computationally expensive, various studies of optimization techniques have been conducted to solve SVM+, such as extended L1-norm SVM [2], L2-norm SVM+ [3] and W-SVM [4].

Although many LUPI algorithms have been proposed in recent literature, the efficient modelling of variational forms of privileged information remains a challenging task, which prevents the wider adoption of this technique in practice. 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, privileged information exists as continuous, ordinal and binary values. An example of continuous privileged information is the tempo of a piece of music which is an essential musical element that can evoke emotions of excitement in the audience. For ordinal privileged information in the middle of Figure 1,

Y.Shu, Q.Li, S.Liu and G.Xu, are with Data Science and Machine Intelligence Lab and the Faculty of Engineering and Information Technology, University of Technology Sydney, Australia (e-mail: Yangyang.Shu@student.uts.edu.au, {Qian.Li, Shaowu.Liu, and guandong.xu}@uts.edu.au)

C.Xu is with the School of Information Technologies, Faculty of Engineering and Information Technologies, University of Sydney, Darlington, NSW 2008, Australia (e-mail: c.xu@sydney.edu.au)



topic/task	music emotion recognition	multiple object recognition	photo aesthetic assessment
implicit cues/PI	tempo	person, motorbike, road, barrier	rule of third
forms of PI	continuous PI: tempo=120 bps	ordinal PI: person>motorbike>road>barrier	binary PI: left photo without PI : 0 right photo with PI : 1

Fig. 1: Examples of different forms of privileged information in regression applications

the implicit importance ranking from photo tags indicates that a photo has good content and information, which can be leveraged for multiple object recognition. For binary privileged information, we provide an example of a photo aesthetic assessment task shown on the right of Figure 1. Certain high-level descriptive attributes, such as the rule of third, explicitly predict possible image cues that can be used to evaluate aesthetics. We indicate whether these attributes exist by denoting them as either 0 or 1. Therefore, considering different forms of PI in training will be of more benefit to learning than using unified privileged information.

In this paper, 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 [5]. 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 [6]. 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 applications: music emotion recognition from songs with the assistance of implicit information about music elements 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; and photo aesthetic assessment enhanced by high-level aesthetic attributes hidden in the photos. For music emotion recognition, we conduct experiments on the benchmark dataset MediaEval and the All Music Guide 1608 database (AMG 1608). For multiple object recognition and photo aesthetic assessment, we conduct experiments on the Pascal VOC 2007 database and the Aesthetics Visual Analysis database (AVA), respectively.

The contributions of this paper are as follows:

- 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.
- Despite their success in certain applications, LUPI algorithms have received little attention in applications with noncontinuous privileged information. In this paper, we make the first attempt to introduce LUPI algorithms to new domains, such as multi-object recognition, in which privileged information exists in non-continuous forms.
- We conduct extensive experiments on four public data sets to study the behaviour and performance of the proposed V-SVR+ framework. Also, we annotate two music data sets (MediaEval 2015 and AMG 1608) with privileged information, which are publicly available at: *https://github.com/GANPerf/Music_PI*

II. Related Works

Privileged information acts similar to a teacher that informs his/her students of helpful comments, comparisons and explanations to improve regression performance. PI is available during training but not testing. Successfully leveraging PI will benefit the learning process of a model. Vapnik and Vashist [1] first proposed learning using privileged information (LUPI) paradigm by upgrading a support vector machine (SVM), proposing the SVM+ method and then extending the SVM+ method (e.g. the mixture model of slacks of LUPI).

Because SVM+ is computationally expensive to train, various studies of optimization techniques and SVM+-based algorithms have been proposed to solve SVM+. Niu et al. [2] proposed an extended l_1 SVM model that uses nonlinear kernels. Instead of the original data domain, the proposed model allows the PI to be explored in a transformed feature space compared to the l_2 SVM+ model. Lapin et al. [4] replicated an SVM+

solution via a weighted SVM. Privileged information is first related to importance weighting, and then, prior knowledge was encoded as expressible privileged features using weights. Finally, they chose weights for the weighted SVMs when the privileged features were unavailable. Li et al. [3] proposed efficiently solving linear and kernel SVM+s using an efficient dual coordinate descent algorithm to solve a new optimization problem that is formulated by absorbing the bias term into the weight vector. For kernel SVM+, they used the l_2-loss based on the ρ-SVM formulation. In addition to using clean training and testing data in SVM+, Li et al. [7] derived a robust SVM+ (R-SVM+) algorithm to study the lower bound of perturbations of both example feature data and privileged feature data based on the SVM+ framework in the LUPI paradigm when potential noise exists in the training and testing data.

Initial work on SVM+ was used for binary classification. Several studies then applied SVM+ to multi-class and multi-label tasks. Wang et al. [8] proposed applying privileged information to learn multi-label classifiers and captured the relationship between PI and the available features using similarity constraints, and the dependencies among multiple labels using ranking constraints. You et al. [9] proposed the privileged multi-label learning (PrML) model to comprehensively explore and exploit the relationships among different labels and explained the hypothesis of a label that is evaluated by itself and the other labels. Yang et al. [10] proposed the use of a two-stream fully convolutional network to use bag-level privileged information (privileged bags) that are available in multi-instance multi-label learning. Ji et al. [11] proposed the multitask multi-class privileged information support vector machines (M^2SVMS) learning paradigm to take full advantage of multitask learning and privileged information. Liu et al. [12] proposed the v-K-SVCR+ method for multi-class classification using privileged information, which solves a one-class classification problem.

In addition to SVM+ classification problems, privileged information is also used in certain regression methods. Cai et al. [13] proposed a regression model called the SVM+MTL-based multi-task learning method. Sarafianos et al. [14] used ratios of anthropometric measurements as privileged information in a regression-based method to estimate height using human metrology. Shu et al. [15] proposed a deep convolutional neural network as a rating system and used photo-based and photography-based attributes as privileged information to enhance the learning process of the regression model.

Privileged information has also been used in various fields, such as image categorization [16] [17], facial expression recognition [18], domain adaptation [19] [20] and deep learning [21] [22]. However, to our knowledge, few studies have considered the different forms of privileged information with their corresponding paradigms.

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 [23]. 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) [24]. 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 [25]. 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 [26]. 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 [27].

To summarize, privileged information is ubiquitous in multiple areas and also typically exists for nearly any machine learning problem.

III. Method

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 2.

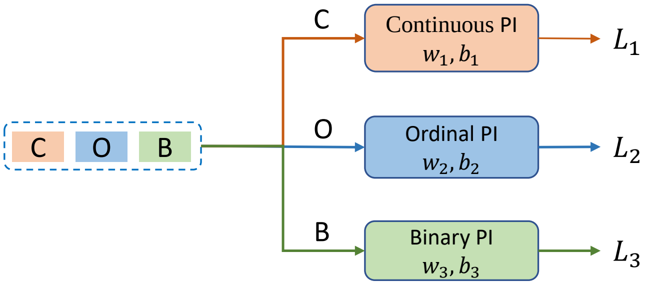


Fig. 2: The framework of the proposed V-SVR+ model, consisting of three modules to manage three forms of PI respectively. The input of $\{C, O, B\}$ represents continuous, ordinal and binary forms of privileged information. w_1, b_1, w_2, b_2, w_3, b_3 are their parameters.

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:

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

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 $y = f(x)$.

A. PI with Continuous Information

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

$$\begin{aligned} f(\mathbf{X_i}) &= (\mathbf{w}, \phi(\mathbf{X_i})) + b, \\ f_1^*(\mathbf{X_i^{*(c)}}) &= (\mathbf{w_1^*}, \phi(\mathbf{X_i^{*(c)}})) + b_1^*, \\ f_2^*(\mathbf{X_i^{*(c)}}) &= (\mathbf{w_2^*}, \phi(\mathbf{X_i^{*(c)}})) + b_2^* \end{aligned} \tag{2}$$

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:

$$\begin{aligned} \min_{w,w_1^*,w_2^*} &\sum_{i=1}^{N}\frac{1}{2}(||\mathbf{w_i}||^2 + ||\mathbf{w_{1i}^*}||^2 + ||\mathbf{w_{2i}^*}||^2) \\ &+ \gamma\sum_{i=1}^{N}[(\mathbf{w_{1i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{1i}] \\ &+ \gamma\sum_{i=1}^{N}[(\mathbf{w_{2i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{2i}^*] \\ \text{s.t. } & Y_i - \mathbf{w_i}^T\phi(\mathbf{X}_i) - b \le \epsilon + (\mathbf{w_{1i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{1i}^* \\ & \mathbf{w_i}^T\phi(\mathbf{X}_i) + b - Y_i \le \epsilon + (\mathbf{w_{2i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{2i}^* \\ & [(\mathbf{w_{1i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{1i}^*] \ge 0 \\ & [(\mathbf{w_{2i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{2i}^*] \ge 0,\ \forall i \end{aligned} \tag{3}$$

To solve this convex optimization problem, we apply the Lagrangian multiplier method [5] to simplify the formulation and improve computational efficiency. The Lagrangian function is shown as follows:

$$\begin{aligned} \mathcal{L}_1 = &\sum_{i=1}^{N}\frac{1}{2}(||\mathbf{w_i}||^2 + ||\mathbf{w_{1i}^*}||^2 + ||\mathbf{w_{2i}^*}||^2) \\ &+ \gamma\sum_{i=1}^{N}[(\mathbf{w_{1i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{1i}^*] + \gamma\sum_{i=1}^{N}[(\mathbf{w_{2i}^*}, \phi(\mathbf{X_i^{*(c)}})) + b_{2i}^*] \\ &+ \sum_{i=1}^{N}\alpha_i(Y_i - f(\mathbf{X_i}) - \epsilon - f_1^*(\mathbf{X_i^{*(c)}})) + \sum_{i=1}^{N}\beta_i(0 - f_1^*(\mathbf{X_i^{*(c)}})) \\ &+ \sum_{i=1}^{N}\alpha_i^*(f(\mathbf{X_i}) - Y_i - \epsilon - f_2^*(\mathbf{X_i^{*(c)}})) + \sum_{i=1}^{N}\beta_i^*(0 - f_2^*(\mathbf{X_i^{*(c)}})) \end{aligned} \tag{4}$$

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

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

$$\begin{aligned} \min_{\alpha,\alpha^*,\beta,\beta^*} &\sum_{i=1}^{N}\gamma_i(\alpha_i - \alpha_i^*) + \epsilon(\alpha_i + \alpha_i^*) \\ &+ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\mathbf{X_i}, \mathbf{X_j}) \\ &+ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^* + \beta_i^* - \gamma)(\alpha_j^* + \beta_j^* - \gamma)K^*(\mathbf{X_i}^{*(c)}, \mathbf{X_j}^{*(c)}) \\ &+ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i + \beta_i - \gamma)(\alpha_j + \beta_j - \gamma)K^*(\mathbf{X_i}^{*(c)}, \mathbf{X_j}^{*(c)}) \\ \text{s.t.} &\sum_{i=1}^{N}(\alpha_i^* - \alpha_i) = 0, \sum_{i=1}^{N}(\alpha_i^* + \beta_i^* - \gamma) = 0 \\ &\sum_{i=1}^{N}(\alpha_i + \beta_i - \gamma) = 0 \\ &\alpha_i^*, \alpha_i, \beta_i^*, \beta_i \ge 0\ i = 1, ..., N. \end{aligned} \tag{5}$$

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 [28] as follows:

$$\begin{aligned} w &= \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)\mathbf{X}_i, \\ b &= \frac{1}{|S| + |S'|}\left[\sum_{s\in S}(\mathbf{Y_s} - \epsilon - w\mathbf{X_s}) + \sum_{s\in S'}(\mathbf{Y_s} + \epsilon - w\mathbf{X_s})\right], \end{aligned} \tag{6}$$

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

$$\begin{aligned} f_1^*(\mathbf{X}) &= (w_1^*, \phi(\mathbf{X}^{*(c)})) + b_1^* \\ &= \sum_{i=1}^{N}(\alpha_i + \beta_i - \gamma)K^*(\mathbf{X_i}^{*(c)}, \mathbf{X}^{*(c)}) + b_1^*, \\ f_2^*(\mathbf{X}) &= (w_2^*, \phi(\mathbf{X}^{*(c)})) + b_2^* \\ &= \sum_{i=1}^{N}(\alpha_i^* + \beta_i^* - \gamma)K^*(\mathbf{X_i}^{*(c)}, \mathbf{X}^{*(c)}) + b_2^* \end{aligned} \tag{7}$$

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

$$f(\mathbf{X}) = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)K(\mathbf{X_i}, \mathbf{X}) + b. \tag{8}$$

We use the conditional gradient method to solve Eq. 5. The obtained parameters w and b as w_1 and b_1 in the first module of the proposed framework as shown in Figure 2.

B. PI with Ordinal 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: $\mathbf{E} = \{(i,j)|\mathbf{X}_i^{*(o)} \leq \mathbf{X}_j^{*(o)}\, or\, \mathbf{X}_i^{*(o)} > \mathbf{X}_j^{*(o)}\}$, 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:

$$\begin{aligned} f(\mathbf{X}_i) &= (\mathbf{w}, \phi(\mathbf{X}_i)) + b, \\ f^*(\mathbf{X}_i^{*(o)}) &= (\mathbf{w}^*, \phi(\mathbf{X}_i^{*(o)})) + b^*, \end{aligned} \tag{9}$$

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:

$$\begin{aligned} |Y_i - f(\mathbf{X}_i)| &\leq \epsilon + \eta_i, \\ |Y_i - f^*(\mathbf{X}_i^{*(o)})| &\leq \epsilon + \eta_i, \end{aligned} \tag{10}$$

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:

$$f^*(\mathbf{X^{*(o)}}_i) - f^*(\mathbf{X^{*(o)}}_j) \geq 1 - \xi_{ij}^*, \tag{11}$$

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:

$$\begin{aligned} &\min_{\theta,\eta,\xi} \frac{1}{2}(||\mathbf{w}||^2 + ||\mathbf{w}^*||^2) + \gamma_1 \sum_{n=1}^{N}\sum_{k=1}^{N}(\eta_k^{(n)+} + \eta_k^{(n)-}) + \\ &\quad \gamma_2 \sum_{n=1}^{N} \sum_{(i,j)\in\mathbf{E}} \xi_{ij}^{*(n)} \\ &\text{s.t. } \mathbf{w}^T\mathbf{X}_k^{(n)} + \mathbf{w^*}^T\mathbf{X^{*(o)}}_k^{(n)} + b + b^* - Y_k^{(n)} \leq \epsilon + \eta_k^{(n)+} \\ &\quad Y_k^{(n)} - \mathbf{w}^T\mathbf{X}_k^{(n)} - \mathbf{w^*}^T\mathbf{X^{*(o)}}_k^{(n)} - b - b^* \leq \epsilon + \eta_k^{(n)-} \\ &\quad \mathbf{w^*}^T(\mathbf{X^{*(o)}}_i^{(n)} - \mathbf{X^{*(o)}}_j^{(n)}) \geq 1 - \alpha_{ij}\xi_{ij}^{*(n)} \\ &\quad \eta_k^{(n)+}, \eta_k^{(n)-}, \xi_{ij}^{*(n)} \geq 0 \\ &\quad (i,j) \in \mathbf{E}, n = 1,...,N \end{aligned} \tag{12}$$

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 [6]. The complexity for solving Eq (12) is $O(n^2)$, where n is the number of samples.

Because ADMM cannot be directly applied to Eq. (12), we first transform Eq. (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. (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. (12) as

$$\begin{aligned} &\min_{\theta,t} \frac{1}{2}\theta^T\Lambda\theta + \mu^T t \\ &\text{s.t. } A\theta - t = c \end{aligned} \tag{13}$$

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,

$$A = \begin{bmatrix} \mathbf{X} & \mathbf{X^{*(o)}} & \mathbf{1} & \mathbf{1} \\ -\mathbf{X} & -\mathbf{X^{*(o)}} & -\mathbf{1} & -\mathbf{1} \\ \mathbf{0} & -\mathbf{X}_E & \mathbf{0} & \mathbf{0} \end{bmatrix} c = \begin{bmatrix} \epsilon\mathbf{1} + \mathbf{Y} \\ \epsilon\mathbf{1} - \mathbf{Y} \\ -\mathbf{1} \end{bmatrix} \tag{14}$$

where $\mathbf{X}_E$ is a matrix whose rows are the difference between two PIs whose indices belong to set $\mathbf{E}$. $\mathbf{1}$ and $\mathbf{0}$ are vectors with the proper dimensions containing all 1s and 0s, respectively. $\mathbf{Y}$ 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.

$$\begin{aligned} \mathcal{L}_2 = L_\rho(\theta, t, v) =& \frac{1}{2}\theta^T\Lambda\theta + \mu^T t + v^T(A\theta - t - c) \\ &+ \frac{\rho}{2}||A\theta - t - c||_2^2 \end{aligned} \tag{15}$$

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

$$\frac{\partial L_\rho(\theta, t, v)}{\partial \theta} = \theta\Lambda + v^T A + \rho A^T ||A\theta - t - c||_1 \tag{16}$$

$$\frac{\partial L_\rho(\theta, t, v)}{\partial t} = \mu^T - v^T - \rho A||A\theta - t - c||_1 \tag{17}$$

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

$$\theta^{k+1} := [\frac{1}{\rho}\Lambda + A^T A]^{-1} A^T (t^k - \frac{1}{\rho}v^k + c) \tag{18}$$

$$t_i^{k+1} := \frac{1}{\rho}v^k + A\theta^{k+1} - \frac{1}{\rho}\mu - c \tag{19}$$

$$v^{k+1} := v^k + \rho(A\theta^{k+1} - t^{k+1} - c) \tag{20}$$

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. 15 is also convex. Thus, Eq. 18 and Eq. 19 can obtain the unique optimal solution and these updates are convergent [6]. 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

$$\mathbf{t}_i^{k+1} = T_\lambda(z_i) = \begin{cases} z_i - \lambda, & \text{if} \quad z_i > \lambda \\ 0, & \text{if} \quad |z_i| \leq \lambda \\ z_i + \lambda, & \text{if} \quad z_i < -\lambda \end{cases} \tag{21}$$

where $\lambda \geq 0$ is the soft threshold operator and i is the i^{th} 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 2.

C. PI with Binary 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:

$$\begin{aligned} p(\hat{Y} = S^+|X^{*(b)} = 1) > p(\hat{Y} = S^-|X^{*(b)} = 1) \\ p(\hat{Y} = S^-|X^{*(b)} = 0) > p(\hat{Y} = S^+|X^{*(b)} = 0) \end{aligned} \tag{22}$$

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. 22 is encoded as follows:

$$\begin{aligned} &\ell_{im}(\mathbf{X_i}, X_{im}^{*(b)}, \hat{Y}_i) = X_{im}^{*(b)} * [p(\hat{Y}_i = S^-|X_{im}^{*(b)} = 1) \\ &- p(\hat{Y}_i = S^+|X_{im}^{*(b)} = 1)]_+ + (1 - X_{im}^{*(b)}) * [p(\hat{Y}_i = S^+|X_{im}^{*(b)} = 0) \\ &- p(\hat{Y}_i = S^-|X_{im}^{*(b)} = 0)]_+ \\ &= X_{im}^{*(b)} * [1 - 2 * p(\hat{Y}_i = S^+|X_{im}^{*(b)} = 1)]_+ \\ &+ (1 - X_{im}^{*(b)}) * [2 * p(\hat{Y}_i = S^+|X_{im}^{*(b)} = 0) - 1]_+ \end{aligned} \tag{23}$$

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

The objective function of binary PI is as follows:

$$\begin{aligned} \mathcal{L}_3 =& \frac{1}{2}\mathbf{w}^T\mathbf{w} + c_1 \sum_{i=1}^{N} \ell_\epsilon(f(\mathbf{X_i}, \mathbf{w}) - Y_i) \\ &+ c_2 \sum_{i=1}^{N} \sum_{m=1}^{d^{*(b)}} \ell_{im}(\mathbf{X_i}, X_{im}^{*(b)}, \hat{Y}_i) \end{aligned} \tag{24}$$

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:

$$\begin{aligned} p(\hat{Y} = S^+|X^{*(b)}) = \sigma(f(X, w)) \\ p(\hat{Y} = S^-|X^{*(b)}) = 1 - \sigma(f(X, w)) \end{aligned} \tag{25}$$

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:

$$w^{(t+1)} = w^{(t)} - \eta^{(t)} \frac{\partial \mathcal{L}_3}{\partial w} \tag{26}$$

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:

$$\begin{aligned} \frac{\partial \mathcal{L}_3}{\partial \mathbf{w}} =& \mathbf{w} + c_1 \sum_{i=1}^{N} \frac{\partial \ell_\epsilon(f(\mathbf{X_i}, \mathbf{w}) - Y_i)}{\partial \mathbf{w}} + \\ & c_2 \sum_{i=1}^{N} \sum_{m=1}^{d^{*(b)}} \frac{\partial \ell_{im}(\mathbf{X_i}, X_{im}^{*(b)}, \hat{Y}_i)}{\partial \mathbf{w}} \end{aligned} \tag{27}$$

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

$$\frac{\partial \ell_\epsilon(f(\mathbf{X_i}, \mathbf{w}) - Y_i)}{\partial \mathbf{w}} = \begin{cases} 0, & if \quad |f(\mathbf{X_i}) - Y_i| \leq \epsilon \\ \phi(\mathbf{X_i}), & otherwise. \end{cases} \tag{28}$$

$$\frac{\partial \ell_{im}(\mathbf{X_i}, X_{im}^{*(b)}, \hat{Y}_i)}{\partial \mathbf{w}} = \begin{cases} -2\sigma(f(\mathbf{X_i}, \mathbf{w}))[1 - \sigma(f(\mathbf{X_i}, \mathbf{w}))]\phi(\mathbf{X_i}), \\ if \ X_{im}^{*(b)} = 1 \ and \ 1 - 2\sigma(f(\mathbf{X_i}, \mathbf{w})) \geq 0 \\ 2\sigma(f(\mathbf{X_i}, \mathbf{w}))[1 - \sigma(f(\mathbf{X_i}, \mathbf{w}))]\phi(\mathbf{X_i}), \\ if \ X_{im}^{*(b)} = 0 \ and \ 2\sigma(f(\mathbf{X_i}, \mathbf{w})) - 1 \geq 0 \\ 0, \qquad otherwise. \end{cases} \tag{29}$$

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 2.

IV. Efficiency Analysis of the Proposed V-SVR+ Framework

TABLE I: Comparison of models in time complexity.

	Sample size (n)	Time complexity	Exec.Time (s)
V-SVR+(Continuous)	25,800/1206	$\mathcal{O}(n)$	47/30
V-SVR+(Ordinal)	5011	$\mathcal{O}(n^2)$	180
V-SVR+(Binary)	210,000	$\mathcal{O}(n)$	783
Adv-DCRN [29]	210,000	-	9600
PI-DCNN [30]	210,000	-	6200

In this section, we discuss the time complexity and running time for convergence or to reach the maximum number of iterations of the proposed V-SVR+ framework shown in Table I. We also discuss two current LUPI methods regarding photo aesthetics tasks, where n is the number of training samples. We find that the time complexity of SVR with the continuous PI (V-SVR+, $\lambda_2, \lambda_3 = 0$) model and SVR with the binary PI model (V-SVR+, $\lambda_1, \lambda_2 = 0$) is $\mathcal{O}(n)$. The time complexity of SVR with the ordinal PI model

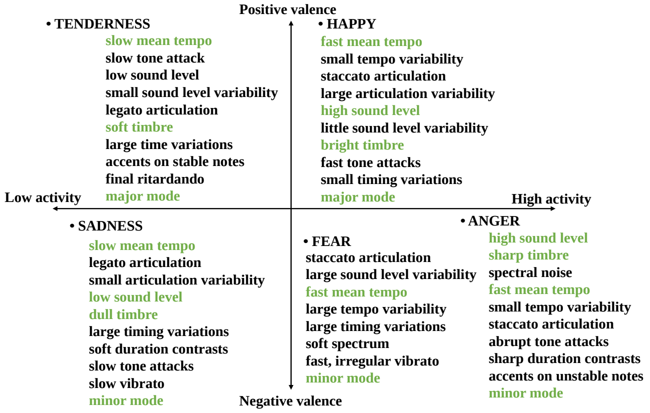


Fig. 3: Musical elements used by composers to communicate emotions to audiences

(V-SVR+, $\lambda_1, \lambda_3 = 0$) is $\mathcal{O}(n^2)$ due to the large number of constraints on ordinal relations. We also determine the approximate execution time when the loss decreases below a given tolerance value or reaches the maximum number of iterations on the four benchmark databases. Then, we find that the V-SVR+ framework converges in a shorter execution time on the aesthetics assessment task compared to the other methods.

V. Experiments

We use two music databases, the `MediaEval` [31] and All Music Guide 1608 databases (`AMG 1608`) [32], and two photo databases, the `Pascal VOC 2007` [33] and the Aesthetics Visual Analysis (`AVA`) databases [34], to conduct the following experiments.

A. Experimental conditions

1) Continuous PI-related music emotion recognition: Music emotion recognition involves the time-continuous estimation of emotion in music, which is typically performed in two dimensions: the arousal space and the valence space [35]. Many musical elements, including tempo, mode, brightness and loudness, are used by composers to affect the audience's experience [36]. In Figure 3, we summarize the relationship between these four musical elements and emotion, and highlight the four primary dimensions. These four musical elements can be extracted by the MIR toolbox [37]. Wang et al. [38] and Shu et al. [30] found that tempo, brightness and loudness have a strong relationship with the arousal space, while mode has a strong relationship with the valence space. These musical elements exist in songs as continuous values when audiences listen to music, however, they are difficult to obtain during testing. Thus, considering their relations with the emotion dimension, we use *tempo, brightness* and *loudness* as privileged information in arousal space and *mode* as privileged information in valence space. The details of privileged information in music emotion recognition are shown in Table II.

The `MediaEval` dataset [1] is split into a training set and a testing set. The training set contains 430 music clips with each clip being 45 s in length, and the testing set contains 58 complete music pieces with an average of 234 s per clip. The `AMG 1608` dataset contains 1608 preview clips of Western songs, which are collected from a popular music stream service called 7digit. We use four-fold cross-validation on this database. For these two databases, we use the features provided by the database promulgator and extract the four musical elements as privileged information using the MIR toolbox [37].

2) Ordinal PI-related multiple object recognition: The goal of multiple object recognition is to detect which instances the image contains, such as cars, people and dog. The implicit cues of importance among the objects in the images are beneficial for multiple object recognition, as shown in Figure 4. We regarded the implicit cues of importance as privileged information obtained from the image absolute tag rank.

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 [39] [40] and 64 dimensional color histogram features [41] [40] in the experiments. Specifically, the 339-dimensional absolute tag rank features used as privileged information during training are provided in [41] where 20 classes as target labels are given with continuous values. Thus, we can use a regression method to finish this classification task.

3) Binary PI-related photo aesthetic assessment: The photo aesthetic assessment task aims to assess photo quality accurately with the assistance of different classifiers and deep models. The aesthetic attributes of photos are typically used to evaluate photo quality.

The `AVA` database contains 250,000 photos collected from a social network [2], and is divided into a training set of 230,000 photos and a testing set of 20,000 photos. The 20,000 photos are selected randomly from the training set as the validation set. In this database, promulgators provide 14 types of aesthetic attributes with a binary value of 0 or 1. The 14 attributes are Soft Focus, Complementary, Light on White, HDR, Photo Grain, Duotones, Shallow DoF, Long Exposure, Motion Blur, Negative Photo, Rule of Thirds, Macro, Silhouettes and Vanishing Point. These attributes are considered to be PI in binary form, which is available in training but not available in testing. Some of these attributes are shown in Figure 5.

The features used in the `AVA` database are extracted by the proposed designed network. Specifically, each photo is first rescaled so that the length of the shorter side is 256. Then, 224×224 patches are cropped randomly from the rescaled photo for data augmentation [42]. Finally, we extract the 2048 dimensional size of the feature representation using the PyTorch and ResNet [43] model from the pre-trained ResNet-152.

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

[2] http://www.dpchallenge.com

TABLE II: The statistics for privileged information in the `MediaEval 2015`. Note that √ denotes a strong dependency between privileged information and arousal/valence. In the valence space, only one element (mode) is related.

ID	PI	Musical dimensions	Description	Arousal	Valence
1	Tempo	Rhythm	Expresses the rhythm and fluency of the music	√	
2	Mode	Tonality	A system of musical tonality		√
3	Brightness	Timbre	A powerful component in constructing the musical piece	√	
4	Loudness	Dynamics	Typically, used by musicians to deliver dynamics in a musical piece	√	

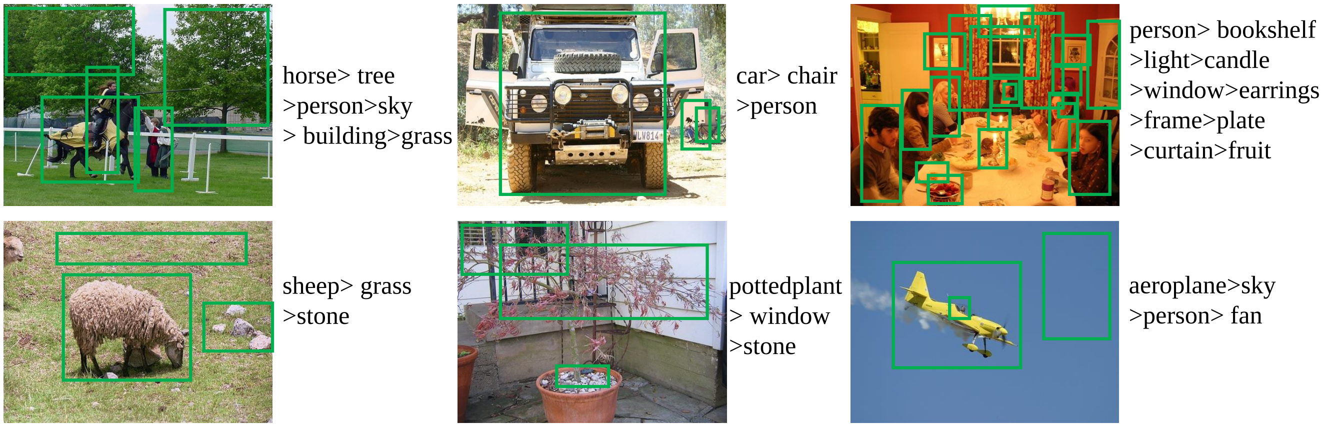


Fig. 4: Example PASCAL images with relative importance of the objects collected on Mechanical Turk.

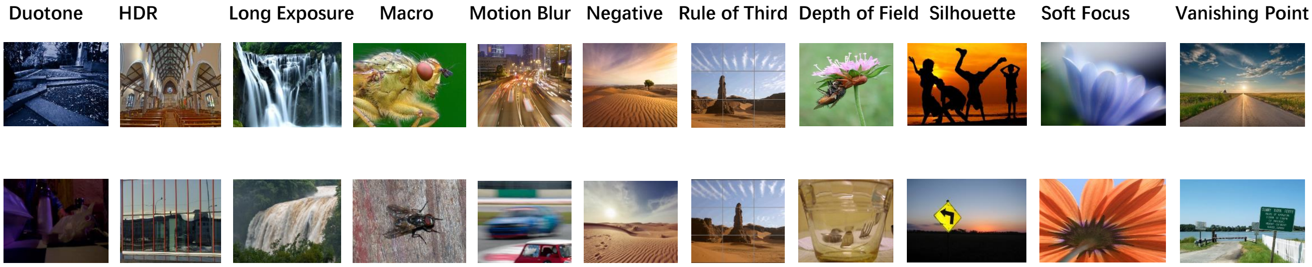


Fig. 5: Different aesthetic qualities w.r.t. different high-level attributes. The top row includes photos with high aesthetic qualities. The bottom row includes photos with low aesthetic qualities.

4) Experimental design: For music emotion recognition with the assistance of implicit information about music elements, we conduct the following experiments in the arousal space: music emotion estimation without any PI (**none**), music emotion estimation with a single PI (**tempo, brightness or loudness**), music emotion estimation with two PIs and music emotion estimation with all PIs. In the valence space, we design the following experiments: music emotion estimation without any PI (**none**) and with mode (**mode**). In the valence space, the mode is the only musical element that affects the valence space. Also, we miss music elements at random with certain probabilities (10%, 20%, 30%, 40%, 50%, 60% and 70%).

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 [40], 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+**).

For the photo aesthetic assessment enhanced by high-level aesthetic attributes, we first conduct experiment on 14 types of privileged information. Then, we use all PI compared to SVR and the deep learning method without PI (deep learning with Euclidean loss only). Specifically, the SVR and V-SVR+ methods use the same features as deep learning at the feature level. The differences are the decision-level where the methods use the **V-SVR+** model, ignore the PI model (**SVR**), and deep learning which uses a fully connected network with Euclidean loss only (**DL**). We also miss 14 kinds of aesthetic attributes at random with probabilities similar to music emotion recognition.

After conducting the contrasting experiment, a statistical t-test is used to analyse whether there are meaningful differences between the model using privileged information and the model that does not use privileged information by considering the p-values shown in Table VI.

5) Performance metrics: The root mean square error (RMSE), average Euclidean distance (AED) and Pearson correlation (R) are used to evaluate the effectiveness of the proposed V-SVR+ framework for the music emotion recognition databases. In multiple object recognition, we use the average precision (AP) as the metric used in the PASCAL challenge [3] [33]. For the photo aesthetic assessment, on the AVA database, we use the Spearman rank-order correlation coefficient (SRCC) and accuracy as the metric evaluation [44]. In these metrics, the smaller the RMSE and AED, the better; the opposite holds for R, AP, SRCC and accuracy.

B. Experiment results and analysis

1) Experimental results with continuous PI with the MediaEval 2015 and AMG 1608 databases: The experimental results for music emotion recognition with continuous privileged information with the MediaEval 2015 database and AMG 1608 database are shown in Table III(a) and Table III(b) respectively. From Table III, we make the following observations:

First, the proposed model with more musical privileged information performs better on music emotion recognition with a lower RMSE and higher R, compared to different combinations of music-related privileged information. Specifically, with the MediaEval 2015 database, the model combining privileged information about tempo and brightness (i.e. t+b model) can improve performance by 0.035 for RMSE and 0.102 for R in the arousal space. Adding loudness to the brightness model (i.e. b+l) can improve RMSE by 0.015 and R by 0.077 in the arousal space. Augmenting the brightness model using tempo and loudness privileged information (i.e. t+b+l) decreases RMSE by 0.049 and increases R by 0.186 respectively. Similar observations and results can be obtained with the AMG 1608 database. Each musical element represents a song using different metrics, and their effects on music emotion recognition are complementary. Therefore, leveraging more privileged information results in better recognition performance.

Second, the models that leverage larger percentages of privileged information achieve better performance than those with smaller percentages. For example, with the MediaEval 2015 database, the proposed method decreases RMSE by 0.01 and 0.068, and increases R by 0.002 and 0.244 for arousal respectively, compared to missing 10% and 70% of proportions. In the valence space, the proposed method decreases RMSE by 0.015 and 0.126, and increases R by 0.021 and 0.096 respectively, compared to missing 10% and 70% proportions. For the AMG 1608 database, we draw similar conclusions. Larger percentages of privileged information involve more underlying knowledge and guidance that can benefit the music emotion recognition task.

2) Experimental results with ordinal PI on the Pascal VOC07 database: With the Pascal VOC 2007 database, the recognition performance of every label is shown in Table IV 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 IV, 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 TV-monitor), 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.

3) Experimental results with binary PI on the AVA database: Table V shows the experimental results of photo aesthetic assessment enhanced by binary privileged information on the AVA database. Binary PI equal to 1 represents photos that contain given attributes, while binary PI equal to 0 represents photos that do not contain given attributes. From Table V, we can make the following observations:

First, compared to the method without PI, the method using single privileged information achieves better performance, which indicates that the hidden attributes in a photo as privileged information are beneficial for aesthetic assessment and regression. Therefore, as expected, the method using combined privileged information has better performance than the method without any PI. For example, combining the 14 types of aesthetic attributes can enhance the performance by 0.152 of SRCC and 16.13% of accuracy. Each aesthetic attribute focuses on the photo from different aspects, for example, ”complementary” describes whether the photo has pairs of colours; ”macro” concentrates on whether the camera uses macro etc. Their effects on photo aesthetic assessment are diverse and complementary. Therefore, the methods leveraging more privileged information result in superior recognition performance.

Second, compared to the method, which misses large percentages of privileged information, smaller missing percentages achieve better performance. For example, for the model that is missing 10% and 70% of the available PI, the proposed method (considers all PI) increases SRCC by 0.001 and 0.065, and increases accuracy by 0.47% and 2.19% respectively. Larger percentages of privileged information involve more underlying knowledge and guidance that can benefit photo aesthetic assessment tasks.

4) Analyses of privileged information: We conduct hypothesis testing to evaluate whether the methods that use privileged information are significantly different from the methods that ignore privileged information. We use the two-sample nonparameter hypothesis (the Wilcoxon test) to evaluate the differences in experimental results that do not follow normal distributions. Table VI shows the experimental results of the p-values of the model that uses privileged information and the model that does not on the four databases. In particular, on the AVA database, we add a deep learning model that only considers Euclidean loss as a method that ignores privileged information for comparison because we use deep learning features at the feature level. From Table VI, we find that all p-values are less than 0.05, which demonstrates that there is a meaningful difference between the proposed model and the

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

TABLE III: Average recognition results on the MediaEval 2015 database and the AMG 1608 database with different missing proportions. "t+b" represents "tempo+brightness", "t+l" represents "tempo+loudness", "b+l" represents "brightness+loudness", "t+b+l" represents "tempo+brightness+loudness".

(a) The results of RMSE and R on MediaEval 2015 database. "-/-" denotes RMSE/R.

miss		0%	10%	20%	30%	40%	50%	60%	70%
Arousal	none	0.280/0.343	0.280/0.343	0.280/0.343	0.280/0.343	0.280/0.343	0.280/0.343	0.280/0.343	0.280/0.343
	tempo	0.242/0.507	0.248/0.492	0.254/0.483	0.259/0.472	0.263/0.454	0.264/0.439	0.268/0.412	0.269/0.393
	brightness	0.247/0.496	0.249/0.483	0.252/0.477	0.258/0.452	0.260/0.437	0.266/0.419	0.269/0.403	0.271/0.396
	loudness	0.252/0.476	0.254/0.471	0.256/0.464	0.261/0.456	0.264/0.448	0.264/0.422	0.266/0.406	0.269/0.388
	t+b	0.212/0.598	0.227/0.544	0.236/0.496	0.241/0.478	0.246/0.443	0.2630.420	0.268/0.413	0.271/0.392
	t+l	0.217/0.616	0.231/0.601	0.242/0.588	0.256/0.543	0.259/0.502	0.261/0.479	0.267/0.452	0.272/**0.440**
	b+l	0.232/0.573	0.238/0.553	0.244/0.537	0.253/0.496	0.259/0.457	0.266/0.448	0.268/0.432	0.270/0.395
	t+b+l	**0.198/0.682**	**0.208/0.646**	**0.225/0.607**	**0.237/0.572**	**0.246/0.543**	**0.251/0.501**	**0.258/0.464**	**0.266**/0.438
Valence	none	0.376/0.016	0.376/0.016	0.376/0.016	0.376/0.016	0.376/0.016	0.376/0.016	0.376/0.016	0.376/0.016
	mode	**0.203/0.128**	**0.218/0.107**	**0.234/0.094**	**0.252/0.086**	**0.276/0.072**	**0.288/0.055**	**0.302/0.048**	**0.329/0.032**

(b) The results of AED and R on AMG 1608 database. "-/-" denotes AED/R.

miss		0%	10%	20%	30%	40%	50%	60%	70%
Arousal	none	0.286/0.802	0.286/0.802	0.286/0.802	0.286/0.802	0.286/0.802	0.286/0.802	0.286/0.802	0.286/0.802
	tempo	0.239/0.870	0.246/0.862	0.249/0.859	0.254/0.851	0.258/0.847	0.263/0.836	**0.266**/0.828	**0.271**/0.816
	brightness	0.241/0.867	0.249/0.860	0.252/0.856	0.258/0.847	0.261/0.842	0.267/0.833	0.272/0.824	0.274/0.812
	loudness	0.244/0.868	0.248/0.858	0.254/0.851	0.259/0.846	0.261/0.837	0.266/0.826	0.274/0.821	0.281/0.813
	t+b	0.237/0.873	0.245/0.867	0.249/0.861	0.253/0.852	0.258/0.843	**0.262**/0.832	0.269/0.828	0.274/0.818
	t+l	0.234/0.877	0.239/0.868	0.244/0.859	0.251/0.851	**0.256**/0.842	0.266/0.838	0.271/0.826	0.276/0.821
	b+l	0.238/0.874	0.242/0.866	0.248/0.863	0.256/**0.855**	0.261/0.847	0.265/0.839	0.272/0.831	0.277/0.825
	t+b+l	**0.221/0.882**	**0.233/0.873**	**0.242/0.864**	**0.253**/0.853	0.259/**0.850**	0.264/**0.842**	0.268/**0.833**	0.272/**0.828**
Valence	none	0.292/0.350	0.292/0.350	0.0.292/0.350	0.292/0.350	0.292/0.350	0.292/0.350	0.292/0.350	0.292/0.350
	mode	**0.234/0.560**	**0.242/0.543**	**0.257/0.504**	**0.262/0.482**	**0.268/0.438**	**0.271/0.425**	**0.278/0.402**	**0.282/0.398**

TABLE IV: 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 [45]	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 [46]	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 [47]	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 [33]	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

model that ignores privileged information.

C. Comparison to related works

We use the MediaEval 2015 and AMG 1608 databases to demonstrate the effectiveness of the proposed method with continuous privileged information. We choose the following eight well-known methods for comparison on the MediaEval 2015 database.

- Multiple linear regression (MLR) is the most common form of linear regression and is typically viewed as a baseline method for music emotion recognition [31].
- RNN [48] uses sequence modelling and prediction smoothing to predict the V-A values in music.
- SVR [49] uses a radial basis kernel function.
- Double-scale support vector regression (DS+SVR) [50] dynamically recognizes music emotion.
- The LSTM-RNN model extracts hidden acoustic and psychoacoustic features from the songs that have been previously shown to be effective for dynamic arousal and valence regression [51].
- Deep BLSTM (DBLSTM) exploits the high context correlation among music feature sequences and sequence information for music emotion recognition [52].
- DNN uses the latest findings in deep learning by stacking convolution layers for music emotion recognition [53].
- The DKLR model uses the domain knowledge of music elements and transfers that knowledge to constraints for music emotion analysis [30].

A comparison of the results of this experiments is shown in Table VII(a). Compared to these methods, we use the simplest features but achieve the best performance, which indicates the vital role of PI in our methods. The proposed method uses continuous PI (i.e. four musical elements) that yield better performance compared to other methods.

Using another music emotion recognition database, AMG 1608, we compare our method with five popular methods shown in Table VII(b). Two of these are the same as

TABLE V: Average estimation results of different attributes on the AVA database with different missing proportions. "-/-" denotes SRCC and accuracy respectively, "total" represents all the aesthetic attributes.

miss	0%	10%	20%	30%	40%	50%	60%	70%
none	0.536/64.30%	0.536/64.30%	0.536/64.30%	0.536/64.30%	0.536/64.30%	0.536/64.30%	0.536/64.30%	0.536/64.30%
Complementary	0.611/71.62%	0.602/70.74%	0.593/69.93%	0.574/67.62%	0.566/67.23%	0.561/66.58%	0.554/66.12%	0.542/65.48%
Duotones	0.597/69.82%	0.591/68.45%	0.578/67.64%	0.565/66.92%	0.548/66.17%	0.541/65.53%	0.539/65.17%	0.537/64.98%
HDR	0.595/69.74%	0.590/68.62%	0.581/67.96%	0.569/67.14%	0.557/66.53%	0.546/65.77%	0.540/65.16%	0.538/64.83%
Photo Grain	0.618/72.36%	0.607/70.88%	0.598/69.43%	0.577/67.66%	0.563/67.19%	0.552/66.54%	0.541/66.11%	0.539/65.39%
Light on White	0.608/70.44%	0.596/69.34%	0.588/68.88%	0.573/67.96%	0.564/67.18%	0.557/66.54%	0.546/66.27%	0.542/65.41%
Long Exposure	0.604/70.12%	0.594/69.78%	0.581/68.83%	0.572/68.12%	0.558/67.47%	0.551/66.52%	0.543/66.23%	0.541/65.47%
Macro	0.612/71.23%	0.602/70.04%	0.591/69.41%	0.583/68.33%	0.575/67.54%	0.558/66.52%	0.544/66.18%	0.538/65.21%
Motion Blur	0.598/69.75%	0.587/68.47%	0.580/67.72%	0.569/67.13%	0.562/66.49%	0.547/65.84%	0.542/65.26%	0.538/64.78%
Negative Photo	0.603/70.18%	0.592/69.84%	0.586/68.53%	0.568/67.42%	0.557/66.49%	0.552/65.72%	0.544/65.18%	0.539/64.96%
Rule of Thirds	0.621/72.36%	0.611/71.44%	0.603/70.65%	0.585/69.14%	0.573/68.37%	0.562/67.28%	0.554/66.54%	0.542/65.52%
Shallow DoF	0.594/69.84%	0.587/69.13%	0.576/68.77%	0.561/67.69%	0.549/66.54%	0.542/65.92%	0.539/65.07%	0.537/64.95%
Silhouettes	0.592/69.12%	0.584/68.76%	0.578/67.91%	0.563/67.02%	0.547/66.88%	0.542/65.93%	0.540/65.26%	0.538/64.88%
Soft Focus	0.607/70.11%	0.597/69.56%	0.582/68.71%	0.577/68.13%	0.564/67.26%	0.549/66.83%	0.544/65.89%	0.540/65.32%
Vanishing Point	0.597/69.53%	0.589/68.56%	0.576/67.92%	0.563/67.14%	0.558/66.58%	0.552/66.06%	0.547/65.67%	0.539/65.16%
total	**0.688/80.43%**	**0.681/79.96%**	**0.672/79.64%**	**0.664/79.38%**	**0.653/79.12%**	**0.641/78.65%**	**0.628/78.36%**	**0.623/78.24%**

TABLE VI: Experimental results on the MediaEval 2015 database, the AMG 1608 database, the Pascal VOC 2007 database, and the AVA database.

Database	Method	RMSE (Arousal)	RMSE (Valence)	R (Arousal)	R (Valence)	Mean AP	SRCC	Accuracy	p-Value in t-test
MediaEval 2015	SVR	0.280	0.376	0.343	0.016	-	-	-	**1.62E-07***(Arousal)
	V-SVR+	**0.198**	**0.203**	**0.682**	**0.128**	-	-	-	**0.0034***(Valence)
AMG 1608	SVR	0.286	0.292	0.802	0.350	-	-	-	**0.0112***(Arousal)
	V-SVR+	**0.221**	**0.234**	**0.882**	**0.560**	-	-	-	**0.0259***(Valence)
Pascal VOC 2007	SVR	-	-	-	-	55.7	-	-	**4.28E-09***
	V-SVR+	-	-	-	-	**65.2**	-	-	
AVA	SVR	-	-	-	-	-	0.536	64.30%	**2.62E-04***
	DL	-	-	-	-	-	0.5210	62.82%	**3.84E-04***
	V-SVR+	-	-	-	-	-	**0.688**	**80.43%**	

*Difference is significant if p-value < 0.05

MediaEval 2015, and the other three models are described below:

- The acoustic emotion Gaussians (AEG) model [54] is a generative model which learns from the emotion annotations of multiple subjects in the valence and arousal spaces.
- CDCC [55] explores the cross-dataset generalizability of music mood regression models in VA spaces with music from different cultures.
- The twin Gaussian process (TGP) [56] is used for structured regression to model the VA spaces of mood.

Compared to other methods, the proposed method achieves the best performance for R in arousal and AED on valence but it also achieves the second-best performance for AED on arousal and R on valence. Unlike the results on the MediaEval 2015 database, the improvements are not significant when compared to the other methods, which may be due to the constraints of the AMG 1608 database where a song is provided with a label and the value of the musical elements represents the entire song rather than the musical bars, Thus, the privileged information has not been used thoroughly.

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 [45] 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 [46] 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 [47] 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 IV. 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 [47] 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.

For the photo aesthetic assessment task on the AVA

TABLE VII: Comparison with related work

(a) Comparison with the related works on the `MediaEval 2015` database

Models	Arousal		Valence	
	RMSE	R	RMSE	R
MLR [31]	0.27	0.36	0.37	0.01
RNN [48]	0.247	0.588	0.365	0.029
SVR [49]	0.255	0.510	0.366	0.022
DS+SVR [50]	0.234	0.61	0.366	0.02
LSTM-RNN [51]	0.242	0.611	0.373	0.0004
DBLSTM [52]	0.239	-	0.318	-
DNN [53]	0.214	-	0.240	-
DKLR [30]	0.234	0.597	0.318	0.044
Ours	**0.198**	**0.628**	**0.203**	**0.128**

(b) Comparison with the related works on the `AMG 1608` database

Models	Arousal		Valence	
	AED	R	AED	R
MLR [31]	0.288	0.806	0.288	0.346
AEG [54]	0.287	0.809	0.287	0.400
CDCC [55]	-	0.854	-	0.435
DKLR [30]	0.240	0.817	0.254	0.374
Regression with TGP [56]	**0.203**	0.808	0.236	**0.661**
Ours	0.221	**0.882**	**0.234**	0.560

(c) Comparison with the related works on the `AVA` database

Models	ρ	Accuracy
RRAC [44]	0.5581	77.3%
DCM [57]	-	78.08%
MTRLCNN [58]	-	79.08%
Adversarial-DCRN (ResNet) [29]	0.6313	-
PI-DCNN [15]	0.6578	76.2%
Ours	**0.688**	**80.43%**

database, we compare the proposed method to some state-of-the-art works that primarily use aesthetic attributes because the proposed method uses aesthetic attributes as PI,. Kong et al. [44], Wang et al. [57], Kao et al. [58], Pan et al. [29] and Shu et al. [15] etc. propose to assess photo aesthetics using aesthetic attributes. Thus, we compare the proposed method with their methods as follows:

- RAPA [44]: a branch is added to predict the aesthetic attributes in the penultimate layer of the original network and the final aesthetic score is given based on the features of the aesthetic attributes and content.
- DCM [57]: Deep Chatterjees Machine (DCM) is tailored to learn attributes through parallel supervised pathways. Then, a high-level synthesis network is trained to associate and transform those attributes into the overall aesthetics rating.
- MTRLCNN [58]: a multi-task framework where the aesthetic assessment problem is the primary task, and the semantic recognition task is critical to addressing this problem.
- Adversarial-DCRN [29] is a multi-task adversarial learning method to learn the aesthetic attributes and aesthetic score simultaneously. The authors of this method designed a rating network as a generator and a discriminator for the rating network output attributes and the score. Then, the generated attributes and score are input to the discriminator.
- PI-DCNN [15]: In the PI-DCNN model for aesthetic assessment, the domain knowledge of the aesthetic attributes is firstly summarized as privileged information, and then a deep convolutional neural network enhanced with privileged information is integrated as a type of loss, replacing the softmax loss.

A comparisons of important results is shown in Table VII(c). In the first three methods, networks must learn the additional branch for high-level features before the aesthetic assessment. Also, some methods, such as the MTRLCNN model only use an attribute that does not thoroughly exploit the aesthetic attributes. Compared to the adversarial-DCRN and PI-DCNN methods, privileged information is learned more easily and absorbed in the support vector regression model. Therefore, the proposed method achieves the best performance.

VI. Conclusion

In this paper, 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 also use three different tasks, music emotion recognition, multiple object recognition, and photo aesthetic assessment to demonstrate the proposed methods. 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. For the photo aesthetic assessment, we consider whether a photo contains high-level aesthetic attributes as binary privileged information in the proposed model. We conducted extensive experiments to demonstrate the superiority of the proposed V-SVR+ framework compared to several other methods with four benchmark databases.

References

[1] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural networks*, vol. 22, no. 5-6, pp. 544–557, 2009.

[2] L. Niu and J. Wu, "Nonlinear l-1 support vector machines for learning using privileged information," in *2012 IEEE 12th International Conference on Data Mining Workshops*. IEEE, 2012, pp. 495–499.

[3] W. Li, D. Dai, M. Tan, D. Xu, and L. Van Gool, "Fast algorithms for linear and kernel svm+," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2258–2266.

[4] M. Lapin, M. Hein, and B. Schiele, "Learning using privileged information: Svm+ and weighted svm," *Neural Networks*, vol. 53, pp. 95–108, 2014.

[5] H. Everett III, "Generalized lagrange multiplier method for solving problems of optimum allocation of resources," *Operations research*, vol. 11, no. 3, pp. 399–417, 1963.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[7] X. Li, B. Du, C. Xu, Y. Zhang, L. Zhang, and D. Tao, "R-svm+: Robust learning with privileged information." in *IJCAI*, 2018, pp. 2411–2417.

[8] S. Wang, S. Chen, T. Chen, and X. Shi, "Learning with privileged information for multi-label classification," *Pattern Recognition*, vol. 81, pp. 60–70, 2018.

[9] S. You, C. Xu, Y. Wang, C. Xu, and D. Tao, "Privileged multi-label learning," *arXiv preprint arXiv:1701.07194*, 2017.

[10] H. Yang, J. Tianyi Zhou, J. Cai, and Y. Soon Ong, "Miml-fcn+: Multi-instance multi-label learning via fully convolutional networks with privileged information," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1577–1585.

[11] Y. Ji, S. Sun, and Y. Lu, "Multitask multiclass privileged information support vector machines," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 2323–2326.

[12] J. Liu, W. Zhu, and P. Zhong, "A new multi-class support vector algorithm based on privileged information," *Journal of Information and Computational Science*, vol. 2, 2013.

[13] F. Cai, "Advanced learning approaches based on svm+ methodology." 2011.

[14] N. Sarafianos, C. Nikou, and I. A. Kakadiaris, "Predicting privileged information for height estimation," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 3115–3120.

[15] Y. Shu, Q. Li, S. Liu, and G. Xu, "Learning with privileged information for photo aesthetic assessment," *Neurocomputing*, 2020.

[16] W. Li, L. Niu, and D. Xu, "Exploiting privileged information from web data for image categorization," in *European Conference on Computer Vision*. Springer, 2014, pp. 437–452.

[17] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, "Image classification by cross-media active learning with privileged information," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2494–2502, 2016.

[18] B. Pan, S. Wang, and B. Xia, "Occluded facial expression recognition enhanced through privileged information," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 566–573.

[19] S. Motiian, "Domain adaptation and privileged information for visual recognition," 2019.

[20] N. Sarafianos, M. Vrigkas, and I. A. Kakadiaris, "Adaptive svm+: Learning with privileged information for domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2637–2644.

[21] J. Lambert, O. Sener, and S. Savarese, "Deep learning under privileged information using heteroscedastic dropout," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8886–8895.

[22] P.-A. Kamienny, K. Arulkumaran, F. Behbahani, W. Boehmer, and S. Whiteson, "Privileged information dropout in reinforcement learning," *arXiv preprint arXiv:2005.09220*, 2020.

[23] G.-J. Qi, C. C. Aggarwal, and T. Huang, "Link prediction across networks by biased cross-network sampling," in *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. IEEE, 2013, pp. 793–804.

[24] J. Wang, Z. Zhao, J. Zhou, H. Wang, B. Cui, and G. Qi, "Recommending flickr groups with social topic model," *Information retrieval*, vol. 15, no. 3-4, pp. 278–295, 2012.

[25] J. Tang, X.-S. Hua, G.-J. Qi, and X. Wu, "Typicality ranking via semi-supervised multiple-instance learning," in *Proceedings of the 15th ACM international conference on Multimedia*, 2007, pp. 297–300.

[26] J. Tang, X.-S. Hua, T. Mei, G.-J. Qi, and X. Wu, "Video annotation based on temporally consistent gaussian random field," *Electronics Letters*, vol. 43, no. 8, pp. 448–449, 2007.

[27] G.-J. Qi, C. C. Aggarwal, and T. S. Huang, "Online community detection in social sensing," in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 617–626.

[28] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," 1998.

[29] B. Pan, S. Wang, and Q. Jiang, "Image aesthetic assessment assisted by attributes through adversarial learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 679–686.

[30] Y. Shu and G. Xu, "Emotion recognition from music enhanced by domain knowledge," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2019, pp. 121–134.

[31] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Emotion in music task at mediaeval 2015," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.

[32] Y.-A. Chen, Y.-H. Yang, J.-C. Wang, and H. Chen, "The amg1608 dataset for music emotion recognition," in *ICASSP 2015*. IEEE, 2015, pp. 693–697.

[33] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[34] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2408–2415.

[35] L. A. Feldman, "Valence focus and arousal focus: Individual differences in the structure of affective experience." *Journal of personality and social psychology*, vol. 69, no. 1, p. 153, 1995.

[36] J. Sloboda, *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press, 2011.

[37] O. Lartillot, "Mirtoolbox 1.3. 4 users manual," *Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Finland*, 2011.

[38] S. Wang, C. Wang, T. Chen, Y. Wang, Y. Shu, and Q. Ji, "Video affective content analysis by exploring domain knowledge," *IEEE Transactions on Affective Computing*, 2019.

[39] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in brain research*, vol. 155, pp. 23–36, 2006.

[40] S. J. Hwang and K. Grauman, "Learning the relative importance of objects from tagged images for retrieval and cross-modal search," *International journal of computer vision*, vol. 100, no. 2, pp. 134–153, 2012.

[41] ——, "Accounting for the relative importance of objects in image retrieval." in *BMVC*, vol. 1, no. 2, 2010, p. 5.

[42] S. Ma, J. Liu, and C. Wen Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4535–4544.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[44] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *European Conference on Computer Vision*. Springer, 2016, pp. 662–679.

[45] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[46] S. J. Hwang and K. Grauman, "Reading between the lines: Object localization using implicit cues from image tags," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 6, pp. 1145–1158, 2011.

[47] N. Dvornik, J. Mairal, and C. Schmid, "On the importance of visual context for data augmentation in scene understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[48] T. Pellegrini and V. Barriere, "Time-continuous estimation of emotion in music with recurrent neural networks," in *MediaEval 2015 Multimedia Benchmark Workshop (MediaEval 2015)*, 2015, pp. pp–1.

[49] M. Chmulik, I. Guoth, M. Malik, and R. Jarina, "Uniza system for the" emotion in music" task at mediaeval 2015." in *MediaEval*, 2015.

[50] H. Xianyu, X. Li, W. Chen, F. Meng, J. Tian, M. Xu, and L. Cai, "Svr based double-scale regression for dynamic emotion prediction in music," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 549–553.

[51] E. Coutinho, G. Trigeorgis, S. Zafeiriou, and B. Schuller, "Automatically estimating emotion in music with deep long-short term memory recurrent neural networks," in *CEUR Workshop Proceedings*, vol. 1436, 2015.

[52] X. Li, H. Xianyu, J. Tian, W. Chen, F. Meng, M. Xu, and L. Cai, "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)*. IEEE, 2016, pp. 544–548.

[53] R. Orjesek, R. Jarina, M. Chmulik, and M. Kuba, "Dnn based music emotion recognition from raw audio signal," in *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 2019, pp. 1–4.

[54] J.-C. Wang, Y.-H. Yang, and H.-M. Wang, "Affective music information retrieval," in *Emotions and Personality in Personalized Services*. Springer, 2016, pp. 227–261.

[55] X. Hu and Y.-H. Yang, "Cross-dataset and cross-cultural music mood prediction: A case on western and chinese pop songs," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 228–240, 2017.

[56] S. Chapaneri and D. Jayaswal, "Structured prediction of music mood with twin gaussian processes," in *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 2017, pp. 647–654.

[57] Z. Wang, D. Liu, S. Chang, F. Dolcos, D. Beck, and T. Huang, "Image aesthetics assessment using deep chatterjee's machine," in *2017*

International Joint Conference on Neural Networks (IJCNN). IEEE, 2017, pp. 941–948.

[58] Y. Kao, R. He, and K. Huang, “Deep aesthetic quality assessment with semantic information,” *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1482–1495, 2017.

Yangyang Shu received his B.S. degree in computer science from Anhui University in 2015, and received his M.S. degree in Computer Science in the University of Science and Technology of China in 2018. Now he is currently pursuing his ph.D degree in Engineering and Information Technology in the University of Technology Sydney, Australia. His research interests cover machine learning, pattern recognition and multimedia.

Qian Li has been a Postdoc Research Fellow at the School of Computer Science, Faculty of Engineering and Information Technology, University of Technology Sydney (UTS). She received her Ph.D. in Computer Science from the Chinese Academy of Science. Her general research interests lie primarily in optimization algorithms, topological data analysis, and causal machine learning. Her papers have been published in the top-tier conferences and journals in the field of machine learning and computer vision.

Chang Xu received the Bachelor of Engineering degree from Tianjin University, Tianjin, China, and the Ph.D. degree from Peking University, Beijing, China. While pursuing his Ph.D. degree, Chang received fellowships from IBM and Baidu. He is a Lecturer of machine learning and computer vision with the School of Information Technologies, The University of Sydney, Sydney, Australia. His research interests lie in machine learning, data mining algorithms and related applications in artificial intelligence and computer vision.

Shaowu Liu received his PhD degree in Computer Science from Deakin University in 2016. Currently, he is a postdoctoral research fellow in School of Computer Science and Advanced Analytics Institute, University of Technology Sydney. His current research interests include User Behavior Analytics, Interpretable Machine Learning, and Representation Learning of Knowledge Graphs.

Guandong Xu is currently a Full Professor with the School of Computer Science, University of Technology Sydney, Ultimo, NSW, Australia. He has published 3 monographs in Springer and CRC press and more than 250 journal articles and conference papers in data science, recommender systems, text mining, and social network analysis. Dr. Xu has served as a Guest Editor for Pattern Recognition, the IEEE Transactions on Computational Social Systems, Journal of Software and Systems, and World Wide Web journal etc. He is Assistant Editor-in-Chief of World Wide Web journal.