Tabular Insights, Visual Impacts: Transferring Expertise from Tables to Images

Jun-Peng Jiang [1 2] Han-Jia Ye [1 2] Leye Wang [3] Yang Yang [4] Yuan Jiang [1 2] De-Chuan Zhan [1 2]

Abstract

Transferring knowledge across diverse data modalities is receiving increasing attention in machine learning. This paper tackles the task of leveraging expert-derived, yet expensive, tabular data to enhance image-based predictions when tabular data is unavailable during inference. The primary challenges stem from the inherent complexity of accurately mapping diverse tabular data to visual contexts, coupled with the necessity to devise distinct strategies for numerical and categorical tabular attributes. We propose CHannel tAbulaR alignment with optiMal tranSport (CHARMS), which establishes an alignment between image channels and tabular attributes, enabling selective knowledge transfer that is pertinent to visual features. Specifically, CHARMS measures similarity distributions across modalities to effectively differentiate and transfer relevant tabular features, with a focus on morphological characteristics, enhancing the capabilities of visual classifiers. By maximizing the mutual information between image channels and tabular features, knowledge from both numerical and categorical tabular attributes are extracted. Experimental results demonstrate that CHARMS not only enhances the performance of image classifiers but also improves their interpretability by effectively utilizing tabular knowledge.

1. Introduction

Data in modern machine learning applications can take various forms, including images, text, video, and audio, providing rich and diverse sources of information. Multimodal learning seeks to fuse information from these different modalities (Ngiam et al., 2011; Ye et al., 2016; Ramachandram & Taylor, 2017; Baltrušaitis et al., 2018; Yang et al., 2020), has demonstrated enhanced model accuracy and comprehensiveness across several domains such as recommender systems (Huang et al., 2019; Salah et al., 2020; Baltescu et al., 2022), healthcare (Zhang et al., 2022; Han et al., 2022), and visual question answering (Li et al., 2019; Zheng et al., 2020; Jing et al., 2020).

Despite its potential, different modalities not only contribute distinctively but also vary significantly in their acquisition costs (Zhou, 2018). For example, in the healthcare field, acquiring medical images relies on specialized equipment, while the extraction of detailed and accurate diagnoses requires expert medical knowledge, often a more challenging and costly endeavor. A practical solution involves leveraging multiple modalities during the training phase to facilitate the transfer of expert knowledge from one modality to another, subsequently enhancing the performance of *single-modality models during the inference phase* (Karpathy & Fei-Fei, 2015; Wang et al., 2016; Radford et al., 2021).

Tabular data, characterized by its structured format of rows and columns (McKinney et al., 2010), often encapsulates expert knowledge that is crucial yet underutilized in image-based machine learning tasks. Continuing with the healthcare example, a doctor's diagnosis, typically recorded in tabular format, can provide critical insights, such as specific annotations in MRI images essential for accurate interpretation. Since the structured nature of tabular data significantly differs from unstructured data like images, existing crossmodal transfer methods are unsuitable for tabular data (Kimball & Ross, 2011; Shwartz-Ziv & Armon, 2022; Hager et al., 2023). This paper focuses on *leveraging expert-derived, yet expensive, tabular data to enhance image-based predictions when tabular data is unavailable during inference*.

Although the main idea is to incorporate the tabular attributes as auxiliary information when training the visual model, there are several challenges when making the learned visual embeddings aligned with the tabular data due to the

[1] School of Artificial Intelligence, Nanjing University, Nanjing, China. [2] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China. [3] Key Lab of High Confidence Software Technologies (Peking University), Ministry of Education & School of Computer Science, Peking University, Beijing, China. [4] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. Correspondence to: Han-Jia Ye <yehj@lamda.nju.edu.cn>, Yang Yang <yyang@njust.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

heterogeneity between these two modalities. Not all the tabular attributes are relevant to the corresponding image. For example, in a pet adoption scenario, the tabular data contains not only the type of the pet but also information such as whether the pet is vaccinated or not. Transferring irrelevant information from the tabular data to the image model can create challenges and hinder the learning process of the image model. We expect that by identifying "which subset of attributes to transfer" and transferring the selected tabular knowledge to the visual model, the visual model can learn more accurate information with rich human expert guidance. Moreover, the tabular data has both categorical and numerical features, which usually require diverse processing strategies. A knowledge transfer approach should be able to deal with these two features and keep the interpretability of the model.

To overcome the aforementioned challenges, we propose CHannel tAbulaR alignment with optiMal tranSport (CHARMS) that selectively aligns tabular data attributes with image channels, which may have different semantics (Zeiler & Fergus, 2014). By maximizing the mutual information between visual predictions and the selected tabular attributes, CHARMS effectively transfers relevant expert knowledge from tabular data to images for two types of attributes.

Specifically, the challenge arises from the inconsistent dimensions between tabular attributes and image channels, making it difficult to align them directly. To overcome this, we utilize sample-wise similarity as an intermedium. Subsequently, we employ the optimal transport algorithm (Caffarelli & McCann, 2010; Bonneel et al., 2011; Ye et al., 2018) to align the two modalities effectively. We strengthen the image channels to ensure they capture the relevant tabular knowledge. By incorporating the tabular data as auxiliary information, we maximize the mutual information between image channels and corresponding tabular attributes. Experiments prove the effectiveness of our CHARMS method and visualization experiments provide evidence that our method successfully transfers expert knowledge from the table into the image model. As a result, the visual model becomes more discriminative and effective. To summarize, our contribution is three-fold:

- We emphasize the importance of knowledge transfer from table to images, as this can lead to improved performance and better understanding when tabular data is missing.
- We propose CHARMS to transfer relevant tabular knowledge to images. It aligns attributes and channels by leveraging optimal transport and utilizes tabular data as auxiliary information during transfer.
- Experimental results demonstrate that CHARMS effectively reuses tabular knowledge to improve the visual classifiers. Moreover, our approach offers insightful explanations of the learned visual embedding space.

2. Related Work

Multimodal learning. Data of different modalities, such as image, video, audio, and text, usually overlap in some content, while some information is complementary. Multimodal learning aims to leverage the information in different modalities to learn a better representation and improve the performance for different scenarios. An important task in multimodal learning is multimodal fusion. Previous work used BERT (Su et al., 2019; Li et al., 2020a) or co-attention (Li et al., 2019; Tan & Bansal, 2019) to fuse different modal information. Subsequently, some large models (Li et al., 2021; Jia et al., 2021; Li et al., 2022) were created to align the information of different modalities in terms of their semantic relationships using contrastive learning approach (Tsai et al., 2018). Different pre-training approaches have also been extensively studied (Liang et al., 2020; Huang et al., 2021; Yao et al., 2021; Bao et al., 2022).

Crossmodal transfer. The modality fusion approach directly depends on the integrity of the data from different modalities. However, the reality is often that we do not have access to the data of all modalities. Therefore, another direction of multimodal learning is to construct robust models to cope with missing modalities or crossmodal transfer (Yang et al., 2024). For example, knowledge in missing modalities can be complemented using autoencoders or generative adversarial approaches (Cai et al., 2018; Li et al., 2020b; Pan et al., 2021). Wang et al. (2020) proposed a framework based on knowledge distillation, utilizing the supplementary information from all modalities, and avoiding imputation and noise associated with it. Ma et al. (2021) improves the robustness of Transformer models by automatically searching for an optimal fusion strategy regarding input data. Hager et al. (2023) proposes the first self-supervised contrastive learning framework that takes advantage of images and tabular data to train unimodal encoders. But most of these approaches consider Vision-Language scenarios, audio or video, which have been well investigated and are not suitable for tabular data due to their structured character and the difference between numerical and categorical variables. Our approach fills the gap of multimodal learning on tabular modality by taking it into account.

Learning with tabular data. The learning of tabular data has become an important research direction in the field of machine learning and data science for a long time. Traditional machine learning methods have been widely used on some tabular data, such as decision trees (Quinlan, 1986), support vector machines (Vapnik, 1999), and random forests (Breiman, 2001). These methods usually rely on pre-processing steps such as manual feature engineering and data cleaning, followed by model training and prediction using supervised learning. With the development of deep learning, tabular modeling approach using deep learning (Huang et al., 2020; Gorishniy et al., 2021; Wang & Sun, 2022) is very appealing because this allows tabular data to be used as input to a single modality and trained end-to-end by gradient optimization, which is competitive with GDBT methods (Chen & Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018). In recent years, more and more approaches for tabular data have been proposed (Arik & Pfister, 2021; Hollmann et al., 2022; Yan et al., 2023; Jeffares et al., 2023). However, tabular data usually contains expert knowledge, such as medical diagnosis information of doctors and seismic waveform information, making it costly to acquire. So we consider such a scenario: expert knowledge from the tabular data is used to guide the learning of the image data during training, with the expectation that good performance can be efficiently obtained even when the tabular data is missing during testing.

3. Preliminaries

In this section, we first introduce the crossmodal transfer task, followed by some existing methods and analyses.

3.1. Transfer Knowledge from Table to Images

In a training set $\mathcal{D} = \{\boldsymbol{x}_i^T, \boldsymbol{x}_i^I, y_i\}_{i=1}^N$, y_i is the label for the ith instance which have two modalities. $\boldsymbol{x}_i^I$ represents the image, and $\boldsymbol{x}_i^T$ represents tabular description. Each image $\boldsymbol{x}_i^I$ has three channels (RGB), and for a tabular data $\boldsymbol{x}_i^T \in \mathbb{R}^D$, each of its D dimensions corresponds to an attribute. There are two kinds of attributes, i.e., the numerical one (such as the "body temperature of a patient") and the categorical one (e.g., the "types of tumors"). Assume there are Y classes in total, and $y_i \in [Y] = \{1, \ldots, Y\}$. We mainly use classification tasks as examples, and our discussions could be extended to regression tasks when $y_i \in \mathbb{R}$.

We aim to construct an image predictor f which maps the input image $\boldsymbol{x}_i^I$ to a certain class. The model f can be decomposed into two parts, i.e., $f(\boldsymbol{x}_i^I) = \boldsymbol{W}^\top \phi(\boldsymbol{x}_i^I)$. The feature extractor $\phi(\cdot)$ transforms a raw image into a C-dimensional vector. In the following of this paper, we denote the dimensions of the visual embedding space as "channels". The linear classifier $\boldsymbol{W} \in \mathbb{R}^{C\times Y}$ predicts the label of an image, and we omit the bias term for discussion simplicity. So as the tabular predictor $g(\boldsymbol{x}_i^T) = \boldsymbol{V}^\top \psi(\boldsymbol{x}_i^T)$, where attribute extractor $\psi(\cdot)$ transforms a tabular data into a E-dimensional vector. During the training phase, we minimize the empirical risk of the model via:

$$\min_f \sum_{i=1}^N \ell(f(\boldsymbol{x}_i^I, y_i \mid \boldsymbol{x}_i^T)) \ . \tag{1}$$

ℓ is the loss function that measures the discrepancy between prediction and ground-truth, such as the cross-entropy loss. "|" indicates that the objective is calculated conditioned on the corresponding tabular part $\boldsymbol{x}_i^T$ of $\boldsymbol{x}_i^I$. Our objective is to transfer relevant tabular information into the image predictor f, which may lead to a more discriminative model. In situations where expert knowledge is not available due to its high collection cost, we expect f can still provide accurate predictions given only the image data $\boldsymbol{x}^I$. Therefore, during the test phase, we measure the performance of f based on its prediction accuracy given any test image $\boldsymbol{x}_i^I$.

3.2. Methods for Crossmodal Transfer

One intuitive idea of transferring the tabular knowledge to the image model is to align the two modalities. We review methods designing the alignment from different perspectives, including parameter-based transfer, embedding-based transfer, and output-based transfer.

Parameter-based transfer. Since the model parameters also contain the learned knowledge, the knowledge transfer can also be conducted from the parameter level. This type of method uses a projection matrix to match the parameters of two modalities. Such as Fixed Model Reuse (FMR) (Yang et al., 2017) optimizes both the visual and tabular models during the training phase and learns an auxiliary mapping from the visual embedding to the tabular data.

$$\mathcal{L} = \sum_{i=1}^N \ell\left(f\left(\boldsymbol{x}_i^I\right) + g\left(\boldsymbol{x}_i^T\right), y_i\right) + \mathcal{L}_{\text{reg}}, \tag{2}$$

where $\mathcal{L}_{\text{reg}} = \frac{1}{2}\left\|\boldsymbol{x}_i^T - \phi(\boldsymbol{x}_i^I)\boldsymbol{U}\right\|_F^2$, $\boldsymbol{U}$ is the linear projection between the tabular features and the embedding of images. FMR removes those connected parts corresponding to features $\boldsymbol{x}^T$ gradually and finally vanishes all related components. However, the method in question transfers all parameters from the tabular data to the image model without incorporating modal alignment and selection, resulting in a rather coarse transfer process.

Embedding-based transfer. The method expects to find a subspace in which the embedding of similar images and tabular data is as close as possible, while the embedding of dissimilar images is as far as possible. For example, Multimodal Contrastive Learning (MMCL) (Hager et al., 2023) proposes the self-supervised contrastive learning framework that takes advantage of images and tabular data to train unimodal encoders:

$$\mathcal{L} = \lambda\ell_{I,T} + (1-\lambda)\ell_{T,I}, \quad \boldsymbol{z}_{j_I} = f_{\phi_I}\left(\phi(\boldsymbol{x}_j^I)\right), \tag{3}$$

$$\ell_{I,T} = -\sum_{j\in\mathcal{N}} \log \frac{\exp\left(\cos\left(\boldsymbol{z}_{j_I}, \boldsymbol{z}_{j_T}\right)\right)}{\sum_{k\in\mathcal{N},k\neq j}\exp\left(\cos\left(\boldsymbol{z}_{j_I}, \boldsymbol{z}_{k_T}\right)\right)}, \tag{4}$$

where embeddings are propagated through separate projection heads f_{ϕ_I} and f_{ϕ_T} and brought into a shared latent space as projections $\boldsymbol{z}_{j_I}, \boldsymbol{z}_{j_T}$. $\ell_{I,T}$ is calculated analagously. $\cos$ means cosine similarity. $\mathcal{N}$ denotes all samples in a

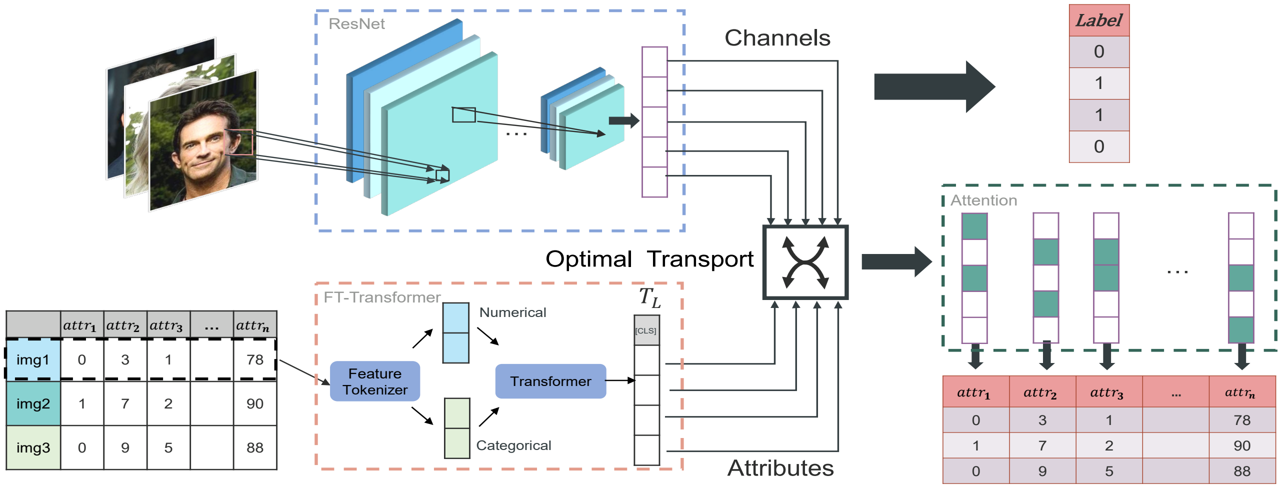


Figure 1: Flowchart of CHARMS method. Our approach combines the learning of image and tabular data, leveraging the specific characteristics of each modality to effectively transfer knowledge from one to the other. We use Optimal Transport methods to match tabular attributes to image channels, effectively learning the correlation attribute of the tabular data with the focused channels as a means of transferring expert knowledge to the images and solving the crossmodal transfer problem.

batch. Then MMCL uses linear probing of frozen networks to evaluate the quality of the learned representations. The embedding-based approach attempts to find a common subspace for alignment but may lose some attribute information in the tabular data when changing the space, potentially ignoring valuable expert knowledge during transfer. Furthermore, not all attributes in the table have corresponding counterparts in the image, making it inappropriate to directly align the entire tabular representation with the image.

Output-based transfer. To transfer knowledge from tabular data to image models, we aim to ensure that the predictions of the image model f and the tabular model g are aligned, so that f is learned to mimic the expert's predictions. The tabular model g such as LightGBM (Ke et al., 2017) is learned on the tabular part of the training set at first. Then in addition to minimizing the discrepancy with labels, an additional distillation term between the predictions of f and g is considered (Hinton et al., 2015):

$$\mathcal{L} = \sum_{i=1}^{N} \left((1-\lambda)\ell(\boldsymbol{x}_i^I, y) + \lambda \mathcal{L}_{\text{KD}}(f(\boldsymbol{x}_i^I), g(\boldsymbol{x}_i^T)) \right). \quad (5)$$

$\mathcal{L}_{\text{KD}}$ measures the similarity between the predictions of two models, e.g., the Kullback-Leibler (KL) divergence. We usually denote g as the "teacher", which transfers the knowledge to the "student" f by aligning its output with g. In Modality Focus Hypothesis (MFH) (Xue et al., 2022), the important attributes from the tabular datasets are selected at first via the off-the-shelf methods. Then the $\mathcal{L}_{\text{KD}}$ is equipped with such modality general decisive information and the distillation is implemented based on the predictions over a subset of the tabular attributes. Such subset comes out to be image-independent, thus the information transfer between the two modalities is also difficult to guarantee.

In summary, the parameter-based approach directly transfers the complete knowledge from the table to the image model, which can potentially overwhelm the image model. Similarly, the embedding-based approach overlooks the fact that different attributes in the table may require distinct processing. While the output-based approach performs feature selection, it does not consider the image-specific context in this selection process.

4. Seeking Alignment for Knowledge Transfer

To tackle the question of "which subset to transfer", it's essential to investigate the influence of various attributes of tabular data on knowledge transfer. To measure the relevance of table attributes and images, we employ mutual information as an evaluation metric. Subsequently, leveraging the alignment, we execute efficient knowledge transfer. The flowchart is shown in Figure 1.

4.1. Preliminary Experiments

First, we want to explore what kind of impact different attributes have on image data. Mutual information is a measure used in information theory to quantify the level of interdependence between two random variables. In this article, it specifically represents the correlation between the information content of the tabular modality and the image modality. To compute mutual information, we employ the MINE (Belghazi et al., 2018) method. If an image model acquires an image representation that incorporates knowledge transferred from the tabular data, the mutual information between this representation and models trained solely on

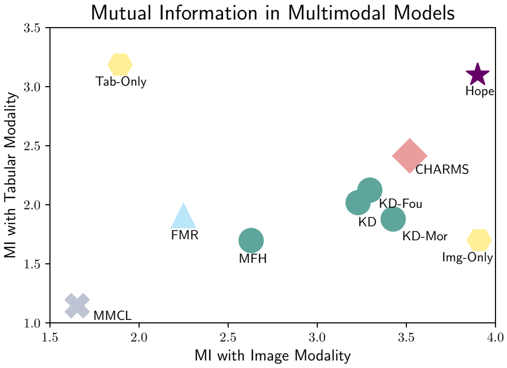


Figure 2: Mutual information with different modality in multimodal models. A good crossmodal transfer model should be able to effectively combine both image and tabular information, resulting in higher mutual information between the two modalities. Ideally, the model should be positioned in the upper right corner. CHARMS is the result of our method, which will be introduced later.

the tabular or image modalities will be high respectively.

To evaluate our approach, we perform experiments on the MFEAT dataset (van Breukelen et al., 1998), which contains two types of tabular data. The first type consists of 76 Fourier coefficients that represent character shapes and lack a direct counterpart in the image. The second type includes 6 morphological features that can be associated with a corresponding part in the image. According to (Artemiou, 2021), there is a positive correlation between mutual information and predictability. Thus, we consider the model trained exclusively on a single modality as the modality optimal model, where mutual information has the largest value. We employ various cross-modal transfer methods to obtain an image model and compute the corresponding image representation. By comparing this with the representations obtained from optimal models, we calculate the respective mutual information. The result is shown in Figure 2.

Our experiments indicate that existing methods for transferring tabular knowledge to image models yield low mutual information between the representations and tabular data. This suggests that these methods are not effective at transferring all types of tabular knowledge to the image modality and that feature selection is crucial. To validate this hypothesis, we perform knowledge distillation of the image model using two models trained on different parts of the tabular data. We find that morphological features in the tabular data can effectively promote image information, while other non-morphological features can make the tabular information more comprehensive.

These results highlight the importance of the selection of different tabular attributes and their relationship with the image modality. Similarly, different channels in image model have different semantics (Zeiler & Fergus, 2014). Through the experiments, we can observe that images and tables exhibit heterogeneity, with not all table attributes being visually apparent in images. Additionally, we can enhance the correlation between the two modalities by focusing on mutual information as a means of facilitating knowledge transfer. Based on these findings, we propose our method for transferring knowledge between modalities, which takes into account the specific characteristics of each modality and transfers expert knowledge to guide the image model.

4.2. Channel Table Alignment

Building upon the previous findings, it becomes apparent that different attributes have varying effects on the image, and different channels of the image hold distinct semantics. This realization serves as motivation to establish correspondences between these channels and the attributes of the table. To proceed with alignment, we need to vectorize the tabular attributes and image channels. However, the challenge lies in establishing the relationship between each channel and attribute. Consequently, it becomes essential to define a measure of similarity that captures their correspondence.

Extract channel representation. To extract representations of the different channels, we use convolutional neural networks (CNNs). CNNs leverage convolutional filters to scan over the input data and extract local features. By stacking multiple convolutional layers, CNNs can learn increasingly complex and abstract features, allowing us to obtain different channels that capture different aspects of the image. Specifically, the channels of image data $\boldsymbol{x}_i^I$ are defined as $\phi_{-1}(\boldsymbol{x}_i^I) \in \mathbb{R}^{H\times W\times C}$, where C is the number of channels. Prior to average pooling, ϕ_{-1} extracts high-level features, resulting in a shape of $H \times W$. Specifically, to address the issue that different channels of an image may have repeated semantics with some redundancy, we use K-Means clustering to group similar channels together. This allows us to obtain fewer distinct C' channels, each capturing a distinct aspect of the image data.

Extract tabular representation. Currently, modern deep tabular data methods employ tokenization and embedding techniques to construct feature representations for tabular data. We use a neural network to obtain the representation of each attribute of the tabular data. This involves transforming all features, including both categorical and numerical variables, into embeddings. The resulting attributes are defined as $\psi(\boldsymbol{x}^{\boldsymbol{T}}) \in \mathbb{R}^{D\times E}$, where D is the number of attributes and E is the embedding dimension. We assume that the first p attributes are numerical variables $\boldsymbol{x}^T_{\text{num}}$, and the remaining q attributes are categorical variables $\boldsymbol{x}^T_{\text{cat}}$.

Align two modalities. Directly establishing the correlation between two modalities is not feasible due to potential differences in dimensionality and semantic inconsistency between them. If two examples exhibit similarity in their representations on a specific attribute as well as a channel, it implies that the semantics of that table attribute and the image channel are also similar. Specifically, for the i^{th} channel and the j^{th} attribute, considering a total of N samples, we can calculate the sample cosine similarity on the channel $\boldsymbol{S}_i^I$ and on the attribute $\boldsymbol{S}_j^T$ independently. Both $\boldsymbol{S}_i^I$ and $\boldsymbol{S}_j^T$ have a shape of $\mathbb{R}^{N\times N}$. Subsequently, we construct the cost matrix $\boldsymbol{C}$ by evaluating the channel-wise similarity against the attribute-wise similarity, where $\boldsymbol{C}_{ij} = \left\| \boldsymbol{S}_i^I - \boldsymbol{S}_j^T \right\|_2^2$.

Based on the cost matrix, we construct a semantic map through the employment of the optimal transport method (Benamou et al., 2015) to minimize the similarity between samples from different modalities. OT is a mathematical framework for measuring the similarity between probability distributions and finding the optimal way to transport mass from one distribution to another. Then the OT transfer matrix is calculated:

$$\begin{aligned} \hat{\boldsymbol{T}} &= \arg\min_{\hat{\boldsymbol{T}}} \langle \hat{\boldsymbol{T}}, \boldsymbol{C} \rangle_F \\ \text{s.t.} \quad \hat{\boldsymbol{T}}\mathbf{1} &\leq \mathbf{a}, \quad \hat{\boldsymbol{T}}^\top \mathbf{1} \leq \mathbf{b}, \quad \hat{\boldsymbol{T}} \geq 0. \end{aligned} \tag{6}$$

Where $\langle\cdot\rangle_F$ denotes the Frobenius norm. $\mathbf{a}$ and $\mathbf{b}$ are source and target distributions. Here source represents the tabular attribute and target represents the image channel. They are both uniformly distributed. After aligning the distributions of the image and tabular data using optimal transport, we obtain the transfer matrix $\hat{\boldsymbol{T}} \in \mathbb{R}^{D\times C'}$. Based on the clustering results, we can restore the corresponding relationship between the tabular attributes and the original channels of the image as $\boldsymbol{T} \in \mathbb{R}^{D\times C}$. Then the channels and attributes are aligned and relevant features are selected.

4.3. Learning with Auxiliary Information

By utilizing OT, we successfully address the feature selection problem by aligning channels with attributes. Building upon the findings from previous experiments in section 4.1, our next objective is to maximize mutual information between two modalities. Since mutual information and predictability exhibit a positive correlation, we aim to enhance the mutual information between the two modalities by partially predicting the corresponding attributes from channels.

Specifically, we employ the transfer matrix $\boldsymbol{T}$ to assign weights to the image channels. This enables us to direct the attention of the relevant tabular attributes towards their corresponding image channels. Utilizing the feature extractor of an existing image network $\phi(\cdot)$, we train a classifier that maps from the image channels to the corresponding attributes. By doing so, we enhance the image network's understanding of the tabular attributes and transfer this knowledge into the image modality. This enables the learned model to effectively handle missing tabular modalities and enhance its performance on complex tasks.

In summary, the loss can be written in the following form

$$\mathcal{L} = \sum_{i=1}^{N} \ell(f(\boldsymbol{x}_i^I), y_i) + \ell(g(\boldsymbol{x}_i^T), y_i) + \mathcal{L}_{i2t}, \tag{7}$$

$$\begin{aligned} \mathcal{L}_{i2t} = &\sum_p \ell_{\text{MSE}}(\boldsymbol{T}_p \cdot \phi(\boldsymbol{x}_i^I), \boldsymbol{x}_{\text{num}p}^T) \\ &+ \sum_q \ell_{\text{CE}}(\boldsymbol{T}_q \cdot \phi(\boldsymbol{x}_i^I), \boldsymbol{x}_{\text{cat}q}^T). \end{aligned} \tag{8}$$

Here, ℓ is the label prediction loss function. ℓ_{CE} is cross entropy loss for categorical attributes and ℓ_{MSE} is mean square error loss for numerical attributes. $\boldsymbol{T}_p$ is the image channels corresponding to the p^{th} numerica attribute, while $\boldsymbol{T}_q$ represents the same for q^{th} categorical attribute.

In our loss function, the first two components correspond to the separate training of the two modalities. The third component $\mathcal{L}_{i2t}$ aims to transfer knowledge from the table attributes to the image channels. The tabular model g is updated to improve the accuracy of representing each tabular attribute in order to calculate $\boldsymbol{S}_T$, thereby facilitating the generation of a more precise transfer matrix for aligning attributes and channels. We update cost matrix every 5 epochs, which ensures that the model learns increasingly accurate channel-attribute correspondences, allowing the tabular data to guide the image data with increasing precision.

To sum up, our approach tackles the challenge of aligning image channels and table attributes by leveraging intersample similarity and OT methods. Subsequently, we aim to transfer the knowledge from the table into the image model by maximizing the mutual information between the two modalities. We handle numerical and categorical variables in the tabular data differently, both in terms of computational representation and final learning processes.

5. Experiments

In this section, we compare CHARMS with crossmodal transfer methods on several datasets. The analysis experiment and ablations verify the effectiveness of our method. Moreover, we visualized the alignment of attributes and channels.

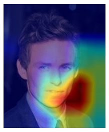
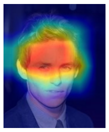
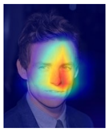
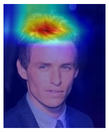
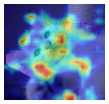
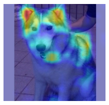
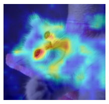
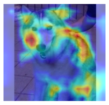
5.1. Experiments and Results

Dataset. Totally six datasets are used in the experiment: **Data Visual Marketing (DVM)** (Huang et al., 2022) is created from 335,562 used car advertisements. The tabular data includes some car parameters such as the number of doors and some advertising data such as the year. Different from (Hager et al., 2023), only the new version DVM dataset is available. Car models with less than 700 samples

Table 1: Comparisons with baseline methods on DVM, SUN, CelebA, Adoption, Pawpularity, and Avito datasets. The first four are classification tasks while the last two are regression tasks. RTDL means the FT-transformer (Gorishniy et al., 2021) model trained on the tabular modality.

	DVM ↑	SUN ↑	CelebA ↑	Adoption ↑	Pawpularity ↓	Avito ↓
LGB	0.9748	0.8501	0.7963	0.4101	20.0720	0.2290
RTDL	0.9682	0.8563	0.7936	0.4107	20.0844	0.2317
Resnet	0.8743	0.8361	0.8146	0.3477	18.6150	0.2512
KD	0.8390	0.8382	0.8118	0.3532	19.0683	0.2499
MFH	–	0.8312	0.7507	0.3041	43.1455	0.2873
FMR	0.8427	0.8347	0.8003	0.3526	19.3517	0.2937
MMCL	0.8203	0.8431	0.8041	0.2981	–	–
CHARMS	**0.9175**	**0.8661**	**0.8220**	**0.3603**	**18.4314**	**0.2495**

Table 2: Visualization by GradCAM. We conducted experiments on CelebA dataset and PetFinder-adoption. The results show that the OT algorithm can indeed align the tabular attributes with the image channels automatically.

Tabular Attribute	5_o_Clock_Shadow	Arched_Eyebrows	Big_Nose	Blond_Hair
Aligned Channel	65, 87, 119, 236…	33, 76, 78, 115, …	50, 224, 258, …	684
Visualization				
Tabular Attribute	Type		Color	
Aligned Channel	399, 413, 414, 521…		400, 412, 425, 448…	
Visualization				

were removed, resulting in 129 target classes, a classification task. **SUNAttribute** (Patterson et al., 2014): We use the table modality in this experiment to help images more accurately predict whether a scene is an open space, which is a binary classification task. **CelebA** (Liu et al., 2015) is the abbreviation of CelebFaces Attribute, meaning celebrity face attribute dataset. It's a large-scale dataset with more than 200K celebrity images, each with 40 attribute annotations. We use Attractive as the label, which is a binary classification task. **PetFinder-adoption** dataset comes from a kaggle competition where the task is to predict the speed at which a pet is adopted, which is a five-class classification task. Tabular data contains information about the pet such as the type and vaccination status. **PetFinder-pawpularity** dataset also comes from a kaggle competition where the task was to predict the popularity of a pet based on that pet's profile and photo. **Avito** is a challenge to predict demand for an online advertisement based on its full description, its context and historical demand for similar ads in similar contexts. The target deal_probability can be any float from zero to one. It's also a regression task.

Evaluation metrics. For classification tasks, we use accuracy to measure the performance. For the regression task, we use root mean square error for performance evaluation.

Implementation Details. In the course of the experiment, we implement CHRAMS with PyTorch and conduct experiments with a single GPU. Moreover, we utilize the grid search to find the hyper-parameters and we choose the best models from the validation set by using early stopping. Specifically, the batch size k is searched in $\{32, 64, 128\}$ and the learning rate is searched in $\{$1e-5, 5e-5, 1e-4, 5e-4, 1e-3, 5e-3$\}$. More details can be seen in Appendix A.

Results. To demonstrate the superiority of CHARMS, we compare it with other popular methods on six datasets as shown in Table 1. The result in the form of mean plus standard deviation are shown in Appendix Table 6. Our results show that CHARMS consistently achieves the best performance on all datasets. In contrast, the baseline methods we evaluated fail to yield significant improvements when compared to direct image training. Some of the baseline methods even exhibit a decline in performance. This outcome can likely be attributed to the limited guidance provided by these methods, which solely rely on tabular data to offer coarse-level guidance to the image model. Consequently, the complex relationships and interactions between modalities are not adequately considered, leading to confusion and subpar results in the image model's learning process.

The MFH approach only learns the KL divergence between the teacher and student networks, which may not be sufficient for handling complex tasks, as evidenced by its poor performance on the DVM 129 classification task. The experiment on the regression task is one of MMCL's limitations according to (Hager et al., 2023).

What is particularly surprising about our approach is that it can outperform the tabular modality on the SUNAttribute dataset. Similarly, on the CelebA and Pawpularity datasets, our approach can improve the performance of the image modality, even though the tabular data is weaker than images. It is possible that our approach can outperform the tabular modality even if it is a strong modality. These findings suggest that we indeed transfer tabular knowledge to images.

Visualization. To verify the effectiveness of OT in matching attributes and channels, we used GradCAM (Selvaraju et al., 2017) to visualize the results of OT, as shown in Table 2. On the CelebA dataset, our model can accurately capture various attributes for the same image. On the PetFinder-adoption dataset, we demonstrate our model's ability to recognize the same attribute across different images.

Our results unequivocally showcase the capability of OT to precisely align image channels with their corresponding tabular attributes, thus affirming the soundness of our approach in seamlessly transferring tabular knowledge into the image model. This finding provides substantial support for the underlying rationale of our approach and emphasizes the criticality of precisely aligning the distributions of diverse modalities to facilitate effective knowledge transfer.

5.2. Experiments Analysis

Comparison for CHARMS and other methods. Throughout the training process, to observe the changes in mutual information, we select ten models at different stages, ranging from the initial training phase to convergence. The outcomes are illustrated in Figure 3. Our findings clearly indicate a consistent and progressive increase in mutual information within CHARMS. This compelling evidence attests to the efficacy of knowledge transfer and substantiates the model's enhanced accuracy and interpretability.

Comparing our approach to the MFH and FMR methods, we observed distinct patterns. Initially, the MFH method, which prioritizes important features, shows higher mutual information with the table. However, as the model increasingly emphasizes image information, the mutual information with the table diminishes. Conversely, the FMR method benefits from a favorable tabular data initialization. Nevertheless, as the table modality is gradually de-emphasized, the mutual information with both the table and image decreases.

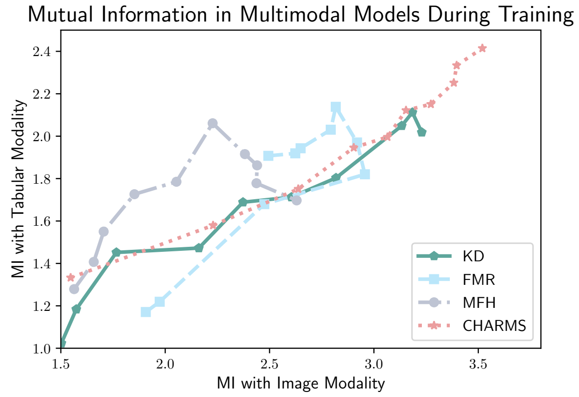


Figure 3: Mutual information during training on MVFEAT dataset. We calculate mutual information from the beginning to the convergence process in order to better understand the training process of each method.

Table 3: Comparison for our CHARMS with CLIP method.

	DVM↑	SUN ↑	CelebA ↑	Adoption ↑
CLIP-LP	0.7619	0.6918	0.7590	0.3047
CLIP-FT	0.8417	0.8333	0.8165	0.2935
CHARMS	**0.9175**	**0.8661**	**0.8220**	**0.3603**

Overall, the visualization of mutual information plays a pivotal role in gaining valuable insights into the learning process of knowledge transfer. It not only enhances the interpretability but also emphasizes the criticality of aligning different modalities and facilitating knowledge transfer.

Comparison with CLIP. CLIP (Radford et al., 2021) is pre-trained on a large amount of text and image pairs, which makes it able to map from text to images. Some previous studies have demonstrated that CLIP is able to transform tabular data to text for classification given the column names (Wang & Sun, 2022; Hegselmann et al., 2023).

In this experiment, we converted the tabular data into text format, such as ”length: 16”. To ensure a fair comparison, we utilize CLIP from (Radford et al., 2021) with the ResNet50 backbone. The CLIP model consists of an image encoder and a textual encoder, and we connect a one-layer linear head for classification or regression after the image encoder. CLIP-LP denotes the scenario where the two encoders are fixed, and only the classification head is trained. CLIP-FT involves fine-tuning the entire CLIP network. By transforming the task into a language-to-vision knowledge transfer, the results are obtained in Table 3.

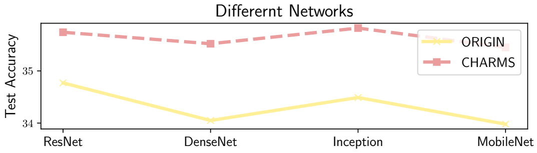


Figure 4: Impact of different network structures on the method on Adoption dataset. It is worth noting that our approach remains unaffected by backbone model.

Table 4: Ablation study on cluster number on SUNAttribute dataset.

cluster	20	40	60	80	100
Accuracy	0.8494	0.8661	0.8494	0.8556	0.8522

From the experiments, we can see that the performance of CLIP is not ideal. This is probably due to the fact that in tabular data, each column holds its own distinct meaning, and directly utilizing it as input to CLIP can lead to the loss of certain information. For instance, on the SUN dataset, the attribute ”wood (not part of a tree)” might not be a highly significant feature. However, when this attribute is converted to text format, its character length tends to be relatively long, which can introduce redundancy in the information.

From another perspective, previous work has pointed out that there is a modality gap in the CLIP’s embedding space (Liang et al., 2022). This gap is caused by a combination of model initialization and contrastive learning optimization. This gap makes the CLIP method fail in our task. For further comparisons on the attention method and CLIP method, please refer to Appendix B.

The ablation study of components in CHARMS. To demonstrate the applicability and robustness of our proposed method, CHARMS, we conducted experiments using different network structures, including Densenet-121, Inception-v1, and MobileNet-v2, in addition to ResNet50. Our results, shown in Figure 4, demonstrate that the performance improvements achieved by our method are consistent across different network structures, highlighting the robustness of our approach. More visualisation and interpretative experiments are provided in Appendix C.

In the CHARMS method, we use the K-Means (Lloyd, 1982; MacQueen, 1967) method to cluster the 2048-dimensional features extracted from ResNet. We discuss the number of clusters on the SUNAttribute dataset, and the results in Table 10 show that the performance of CHARMS is not affected by the number of clusters taken, demonstrating the robustness of the method to hyperparameter choices. This robustness makes the method more flexible and reliable in practical applications, as it does not require excessive hyperparameter tuning or fine-tuning, saving time and effort.

6. Conclusion

In this work, we propose the CHARMS, a new method that automatically transfers relevant tabular knowledge to images. Our method leverages tabular data as auxiliary information during transfer, enabling the transfer of expert knowledge in tabular data to images. Since not all attributes contained in tabular data are relevant to the corresponding image, we utilize optimal transport to align the attributes with channels, strengthening the correlated channels during transfer. Experimental results demonstrate that CHARMS outperforms previous methods in crossmodal transfer and our method enables insightful explanations of the learned visual embedding space with tabular instruction. We hope this work motivates future research on the challenges of multimodal encountered in real-world problems, with a particular focus on tabular data and knowledge transfer.

Acknowledgements

This work is partially supported by National Science and Technology Major Project (2022ZD0114805), NSFC (62250069, 62376118, 62006112, 61921006, 62276131), Collaborative Innovation Center of Novel Software Technology and Industrialization, the Fundamental Research Funds for the Central Universities (No.30922010317).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences in our work, none of which we feel must be specifically highlighted here.

References

Arik, S. Ö. and Pfister, T. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6679–6687, 2021.

Artemiou, A. Using mutual information to measure the predictive power of principal components. In *Festschrift in Honor of R. Dennis Cook: Fifty Years of Contribution to Statistical Science*, pp. 1–16. Springer, 2021.

Baltescu, P., Chen, H., Pancha, N., Zhai, A., Leskovec, J., and Rosenberg, C. Itemsage: Learning product embeddings for shopping recommendations at pinterest. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2703–2711, 2022.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41: 423–443, 2018.

Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O. K., Aggarwal, K., Som, S., Piao, S., and Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37:A1111–A1138, 2015.

Bonneel, N., Van De Panne, M., Paris, S., and Heidrich, W. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pp. 1–12, 2011.

Breiman, L. Random forests. *Machine learning*, 45:5–32, 2001.

Caffarelli, L. A. and McCann, R. J. Free boundaries in optimal transport and monge-ampere obstacle problems. *Annals of Mathematics*, 171:673–730, 2010.

Cai, L., Wang, Z., Gao, H., Shen, D., and Ji, S. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1158–1166, 2018.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34: 18932–18943, 2021.

Hager, P., Menten, M. J., and Rueckert, D. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. *arXiv preprint arXiv:2303.14080*, 2023.

Han, Z., Yang, F., Huang, J., Zhang, C., and Yao, J. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20707–20717, 2022.

Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., and Sontag, D. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pp. 5549–5581, 2023.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

Huang, J., Chen, B., Luo, L., Yue, S., and Ounis, I. Dvm-car: A large-scale automotive dataset for visual marketing research and applications. In *2022 IEEE International Conference on Big Data*, pp. 4140–4147, 2022.

Huang, X., Khetan, A., Cvitkovic, M., and Karnin, Z. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

Huang, Z., Xu, X., Ni, J., Zhu, H., and Wang, C. Multimodal representation learning for recommendation in internet of things. *IEEE Internet of Things Journal*, 6:10675–10685, 2019.

Huang, Z., Zeng, Z., Huang, Y., Liu, B., Fu, D., and Fu, J. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12976–12985, 2021.

Jeffares, A., Liu, T., Crabbé, J., Imrie, F., and van der Schaar, M. Tangos: Regularizing tabular neural networks through gradient orthogonalization and specialization. *arXiv preprint arXiv:2303.05506*, 2023.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916, 2021.

Jing, C., Wu, Y., Zhang, X., Jia, Y., and Wu, Q. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 11181–11188, 2020.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 2017.

Kimball, R. and Ross, M. *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.

Li, G., Duan, N., Fang, Y., Gong, M., and Jiang, D. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11336–11344, 2020a.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34: 9694–9705, 2021.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900, 2022.

Li, L., Du, B., Wang, Y., Qin, L., and Tan, H. Estimation of missing values in heterogeneous traffic data: Application of multimodal deep learning model. *Knowledge-Based Systems*, 194:105592, 2020b.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Y. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625, 2022.

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., et al. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6008–6018, 2020.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, pp. 3730–3738, 2015.

Lloyd, S. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., and Peng, X. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021.

MacQueen, J. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pp. 281–297, 1967.

McKinney, W. et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, pp. 51–56, 2010.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 689–696, 2011.

Pan, Y., Liu, M., Xia, Y., and Shen, D. Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44: 6839–6853, 2021.

Patterson, G., Xu, C., Su, H., and Hays, J. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108:59–81, 2014.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. Catboost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31, 2018.

Quinlan, J. R. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Ramachandram, D. and Taylor, G. W. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34:96–108, 2017.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115: 211–252, 2015.

Salah, A., Truong, Q.-T., and Lauw, H. W. Cornac: A comparative framework for multimodal recommender systems. *The Journal of Machine Learning Research*, 21: 3803–3807, 2020.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shwartz-Ziv, R. and Armon, A. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

Tan, H. and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Tsai, Y.-H. H., Liang, P. P., Zadeh, A., Morency, L.-P., and Salakhutdinov, R. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

van Breukelen, M., Duin, R. P., Tax, D. M., and Den Hartog, J. Handwritten digit recognition by combined classifiers. *Kybernetika*, 34:381–386, 1998.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

Vapnik, V. N. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10:988–999, 1999.

Wang, K., Yin, Q., Wang, W., Wu, S., and Wang, L. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

Wang, Q., Zhan, L., Thompson, P., and Zhou, J. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1828–1838, 2020.

Wang, Z. and Sun, J. Transtab: Learning transferable tabular transformers across tables. *arXiv preprint arXiv:2205.09328*, 2022.

Xue, Z., Gao, Z., Ren, S., and Zhao, H. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*, 2022.

Yan, J., Chen, J., Wu, Y., Chen, D. Z., and Wu, J. T2g-former: organizing tabular features into relation graphs promotes heterogeneous feature interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10720–10728, 2023.

Yang, Y., Zhan, D.-C., Fan, Y., Jiang, Y., and Zhou, Z.-H. Deep learning for fixed model reuse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2831–2837, 2017.

Yang, Y., Zhan, D.-C., Jiang, Y., and Xiong, H. Reliable multi-modal learning: A survey. *Journal of Software*, 32: 1067–1081, 2020.

Yang, Y., Wei, H., Zhu, H., Yu, D., Xiong, H., and Yang, J. Exploiting cross-modal prediction and relation consistency for semisupervised image captioning. *IEEE Transactions on Cybernetics*, 54(2):890–902, 2024.

Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.-S., and Sun, M. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021.

Ye, H.-J., Zhan, D.-C., Li, X., Huang, Z.-C., and Jiang, Y. College student scholarships and subsidies granting: A multi-modal multi-label approach. In *2016 IEEE 16th International Conference on Data Mining*, pp. 559–568, 2016.

Ye, H.-J., Zhan, D.-C., Jiang, Y., and Zhou, Z.-H. Rectify heterogeneous models with semantic mapping. In *International Conference on Machine Learning*, pp. 5630–5639, 2018.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833, 2014.

Zhang, C., Chu, X., Ma, L., Zhu, Y., Wang, Y., Wang, J., and Zhao, J. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2418–2428, 2022.

Zheng, C., Guo, Q., and Kordjamshidi, P. Cross-modality relevance for reasoning on language and vision. *arXiv preprint arXiv:2005.06035*, 2020.

Zhou, Z.-H. A brief introduction to weakly supervised learning. *National Science Review*, 5:44–53, 2018.

A. Experiment Details

A.1. Dataset Details

The datasets used in our experiments are MFEAT (van Breukelen et al., 1998), Data Visual Marketing (DVM) (Huang et al., 2022), SUNAttribute (Patterson et al., 2014), CelebA (Liu et al., 2015), PetFinder-adoption, PetFinder-pawpularity and Avito.

MFEAT. This dataset consists of features of handwritten numerals ('0'–'9') extracted from a collection of Dutch utility maps. 200 patterns per class (for a total of $2,000$ patterns) have been digitized in binary images. These digits are represented in terms of the following six feature sets. We use only 76 fourier coefficients of the character shapes and 6 morphological features for tabular data. The image modality is reconstructed from 240 pixel averages of images from 2×3 windows.

DVM. DVM dataset aims to facilitate business related research and applications in automotive industry such as car appearance design, consumer analytics and sales modeling. The dataset contains car images, model specifications and sales information about 899 car models that have been sold in the UK market over the last 20 years. which comprises two data parts: the image data and the table data. The former contains $1,451$,784 car images that have been deliberately cleaned and organized. While the latter includes six CSV tables that cover the non-visual attributes such as brand, price, sales, etc. Different from MMCL, only the new version DVM dataset is available (Hager et al., 2023). We pair this tabular data with a single random image from each advertisement, yielding a dataset of $70,580$ train pairs, $17,645$ validation pairs, and $88,226$ test pairs. Car models with less than 700 samples were removed, resulting in 129 target classes, classification task. There are total of 13 numerical variables and 3 categorical variables in this dataset. We expect that under the guidance of tabular data, images can learn more knowledge and make classification better.

The DVM dataset utilized in the original paper is an earlier version, and unfortunately, we don't have access to the dataset after the official update. This discrepancy in dataset versions may introduce variations in the data distribution and characteristics. Specifically, all the images are resized to 300x300 resolutions; Segment results are no longer provided directly; Image data of 2019 registered car models is added and the non-visual feature data is updated to 2020.

We follow the steps in (Hager et al., 2023) in Section 4.1 to preprocess the data. In detail, the car models with less than 700 samples were removed, resulting in 129 target classes. This process ensures that the amount of data remain largely consistent with (Hager et al., 2023).

Lastly, to maintain uniformity and facilitate fair comparisons, we employed a fixed batch size of 64 across all methods, whereas the original paper employed a larger 512. Additionally, we conducted MMCL method on our dataset with a batch size of 512. The result was 0.8869/0.9070. This is still somewhat different from the values reported in (Hager et al., 2023) and performs worse than our method 0.9207 with a batch size of 512.

Furthermore, we conducted a comparison of GPU usage with batch size 64. Our method uses 8 GB of memory while theirs uses 20 GB. The results revealed that the MMCL method remains resource-intensive. Conversely, our method achieves superior performance with lower computational costs, further highlighting the efficiency of our approach.

SUNAttribute. SUNAttribute annotates 20 scenes from each of the 717 SUN categories. Each scene has 102 attributes and each attribute will have multiple annotations. For simplicity, we divide each attribute into zero and one and our goal is to predict whether a scene is an open space, which is a binary classification task. The dataset contains $14,340$ images and the corresponding table feature, each attribute of the table feature represents a scene and takes the value of 1 if the attribute is present in the image. we use $8:1:1$ to divide the training set, validation set, and testing set. There are total of 101 categorical variables in this dataset.

CelebA. CelebA is the abbreviation of CelebFaces Attribute, meaning celebrity face attribute dataset, which contains $202,599$ face images of $10,177$ celebrities, each image is well marked with features, including 40 attribute markers such as Big_Nose. We use Attractive as the label, which is a binary classification task. We use $8:1:1$ to divide the training set, validation set, and testing set. There are total of 39 categorical variables in this dataset. We expect to introduce more detailed face information in the table, allowing the image to perform better on downstream tasks.

PetFinder-adoption. Animal adoption rates are strongly correlated to the metadata associated with their online profiles, such as descriptive text and photo characteristics. This dataset comes from a kaggle competition where the task is to predict the speed at which a pet is adopted, which is a five-class classification task. There are total of 10 numerical variables and 14 categorical variables in this dataset. Tabular data contains information about the pet such as the type and vaccination status.

We also use the same division for the dataset.

PetFinder-pawpularity. This dataset also comes from a kaggle competition where the task was to predict the popularity of a pet based on that pet's profile and photo, which is a regression task. Each pet photo is labeled with the value of 1 (Yes) or 0 (No) for each of features. For example, "Face" represents whether the face of the pet in the picture is frontal. There are 12 categorical variables in tabular data.

Avito. Avito, Russia's largest classified advertisements website, is deeply familiar with this problem. Sellers on their platform sometimes feel frustrated with both too little demand (indicating something is wrong with the product or the product listing) or too much demand (indicating a hot item with a good description was underpriced). This dataset is challenging you to predict demand for an online advertisement based on its full description, its context and historical demand for similar ads in similar contexts. The target deal_probability can be any float from zero to one. It's also a regression task. There are total of 2 numerical variables such as and 11 categorical variables such as in this dataset.

Table 5: Introduction to the dataset. Here we introduce image data and tabular data in each dataset, and numerical and categorical variables are introduced separately in the tabular data. An example is given for each dataset.

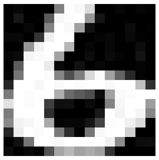
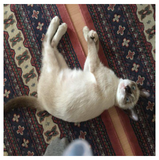
Dataset	Numerical Attribute	Categorical Attribute	Image
MFEAT	Fourier coefficient_1 0.13839	-	
DVM	Length 4865.0	Fuel_type 9	
SUNAttribute	-	Warm 1	
CelebA	-	Big_Nose 0	
PetFinder-adoption	Fee 100	Type 0	
PetFinder-pawpularity	-	Focus 0	
Avito	Price 1290	Category_name 4	

Table 6: Comparisons with baseline methods on DVM, SUN, CelebA, Adoption, Pawpularity, and Avito datasets on five random seeds.

	DVM ↑	SUN ↑	CelebA ↑	Adoption ↑	Pawpularity ↓	Avito ↓
LGB	0.9748±0.0014	0.8501±0.0003	0.7963±0.0005	0.4101±0.0053	20.0720±0.0072	0.2290±0.0011
RTDL	0.9682±0.0018	0.8563±0.0011	0.7936±0.0004	0.4107±0.0048	20.0844±0.0098	0.2317±0.0034
ResNet	0.8743±0.0183	0.8361±0.0144	0.8146±0.0092	0.3477±0.0048	18.6150±1.4559	0.2512±0.0034
KD	0.8390±0.0076	0.8382±0.0063	0.8118±0.0046	0.3532±0.0035	19.0683±1.7642	0.2499±0.0015
MFH	–	0.8312±0.0022	0.7507±0.0034	0.3401±0.0027	43.1455±2.0843	0.2873±0.0047
FMR	0.8427±0.0151	0.8347±0.0119	0.8003±0.0143	0.3526±0.0088	19.3517±1.5837	0.2937±0.0084
MMCL	0.8203±0.0040	0.8431±0.0012	0.8041±0.0017	0.2981±0.0026	–	–
CHARMS	**0.9175±0.0052**	**0.8661±0.0032**	**0.8220±0.0022**	**0.3603±0.0037**	**18.4314**±0.7427	**0.2495±0.0025**

A.2. Training Details

We use ResNet50 with weight pretrained on ImageNet-1k (Russakovsky et al., 2015) as image feature extractor for all methods mentioned in this paper. The classifier is built from an MLP with one hidden layer of size 1024.

For baseline methods, the numerical tabular data fields are standardized using z-score normalization with a mean value of 0 and standard deviation of 1. For our method CHARMS, we use FT-Transformer (Gorishniy et al., 2021) to get the embedding of tabular data, which can process continuous and categorical variables separately.

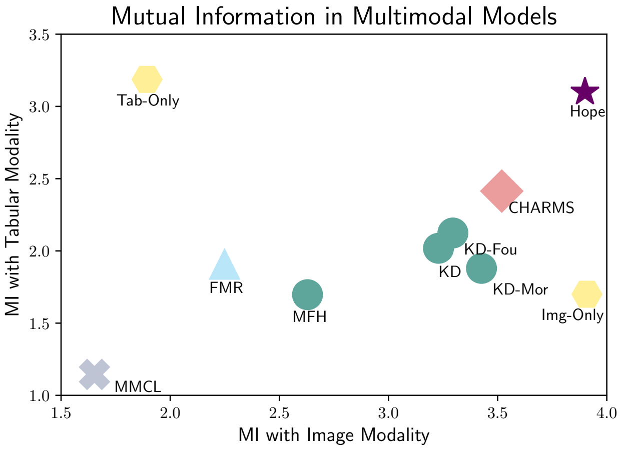
- **KD (Hinton et al., 2015):** For KD method, we search the temperatures in $\{1.0, 2.0, 4.0, 6.0, 8.0\}$ and λ in $\{0.2, 0.4, 0.6, 0.8\}$.
- **KD-Fou:** This means that we use only 76 fourier coefficients of the character shapes features when training the teacher network.
- **KD-Mor:** This means that we use only 6 morphological features when training the teacher network, which can be revealed in images.
- **FMR (Yang et al., 2017):** We set ten percent of the fixed features to be knockdown in each epoch in FMR method. We search the knockdown_num in $\{0.1, 0.2, 0.3, 0.3\}$. The fixed feature classifier is a linear connection between tabular data and the corresponding image.
- **MFH (Xue et al., 2022):** For MFH method, we set modality general decisive information according to the feature ranking algorithm. The number of the features is fifty percent of that for all features.
- **MMCL (Hager et al., 2023):** The same parameters are set for MMCL method according to (Hager et al., 2023). We use the frozen version after pretrain and only train the classifier for downstream task.
- **CHARMS:** For FT-Transformer, the number of Transformer blocks is set to 2. We use the K-Means method to cluster the representations obtained by ResNet50 and $n\_cluster$ is 40. Embedding dimension E is set according to the data distribution. Adam optimizer with weight decay is used to train the models. We choose to update cost matrix every 5 epochs, striking a balance between updating them without stable knowledge and minimizing the computational burden. However, we continuously update ϕ throughout the training process to enhance the representation.

We experiment on five random seeds and the results in the form of mean plus standard deviation are shown in the Table 6.

A.3. Figure Details

We explain some figures in detail.

- For Figure 5, we calculated the amount of information contained in different modality data for different methods with the MINE method (Belghazi et al., 2018). The image data are simple handwritten digits, we process them simply using



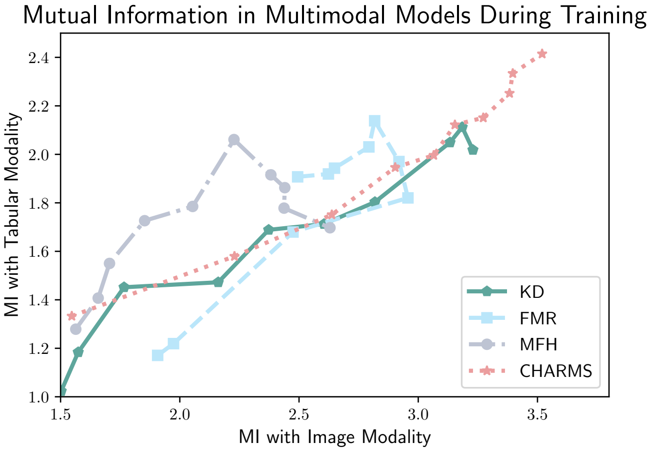


Figure 5: Mutual Information with Different Modality in Multimodal Models. A good model should be able to effectively combine both image and tabular information, resulting in higher mutual information between the two modalities.

Figure 6: Mutual Information During Training on MVFEAT dataset. We calculate mutual information from the beginning to the convergence process in order to better understand the training process of each method.

a two-layer convolutional neural network, followed by a max pooling layer, and a Dropout layer to prevent overfitting. When calculating the mutual information, we use the $mine$ method as the loss function for approximating the mutual information. The network we choose is a three layer MLP with two hidden layers of size 100, the method we choose is $concat$, and the $batch\_size$ is 16.

- For Figure 6, we do not calculate the mutual information change process for the MMCL method because the MMCL method already performs much less well in Figure 8 than the other baseline models. We hypothesize that MMCL maps the tabular and image representations to another space and therefore the mutual information is lower.

- In the ablation study for different nets, we experimentally validated the impact of different neural network as backbone models on our approach. The accuracy in ORIGIN is $\{34.77, 34.05, 34.49, 33.98\}$. The accuracy in out CHARMS is $\{35.74, 35.52, 35.82, 35.45\}$.

A.4. Task Details

The usage of knowledge from table to images could be explained from three aspects:

In our setting, the goal is to transfer knowledge from the tabular data to the image model. Both classification and regression tasks are vital and commonly encountered in our setting, where both of them are investigated in our experiments. For instance, on the Adoption dataset, the pet type and size attributes are crucial for the adoption time classification. Guidance on these features in an image would lead to better learning of the image model. Similarly, on the Pawpularity dataset, the eyes and face attributes have a positive assignment on the regression of the popularity of the pet. Therefore, it makes sense to do knowledge transfer from tabular data to image for both classification and regression tasks.

CHARMS is a general method for both classification and regression tasks, in detail, we use cross entropy loss for classification task and mean square error loss for regression task. We achieved an improved image representation by employing the CHARMS method, which leverages the guidance of tabular data on the image data. Specifically, for the classification task, our approach facilitated the representation with a more discerning distribution over the target categories. On the other hand, the regression task enabled us to learn an image representation that better approximated the target values during prediction. The fact that our method performs well on both tasks underscores its generalizability and effectiveness.

Additionally, our visualization experiments provide further evidence of the effectiveness of our method. These experiments reveal that the attributes and channels selected by our approach are appropriately matched, leading to an enhancement in the performance of the image model. This alignment between the attributes and channels serves as strong evidence that we have successfully transferred the relevant knowledge from the table to the image model.

Table 7: Comparison with Attention method. Here Attention means we directly conduct the attention mechanism on the feature extracted by ϕ and learn an attention mask for all tabular attributes.

	DVM ↑	SUN ↑	CelebA ↑	Adoption ↑	Pawpularity ↓	Avito ↓
Attention	0.4757	0.8550	0.8180	0.3454	18.7401	0.2544
CHARMS	**0.9175**	**0.8661**	**0.8220**	**0.3603**	**18.4314**	**0.2495**

In summary, our approach demonstrates its versatility by excelling in both classification and regression tasks, showcasing its ability to enhance image representations using guidance from tabular data.

B. Analysis on Our CHARMS Method

B.1. Comparison with attention method

Our method employs the transfer matrix obtained by OT to weigh the images, with the weights of the corresponding channels raised to learn the tabular attributes. An alternative approach is to use the attention method to weigh the image channels differently and learn each tabular attribute separately, which is a more intuitive approach:

$$\phi(\boldsymbol{x}^T)_{att} = \mathcal{T}(\phi(\boldsymbol{x}^T)) \cdot \phi(\boldsymbol{x}^T) \tag{9}$$

where $\mathcal{T}$ is a two layer MLP that first downscales the image representation obtained by ϕ before rescaling it to its original dimension, thereby weighting the different channels of the image.

In contrast to our method CHARMS, this method assigns a weight to each input element so that the model can pay more attention to those input elements that are more important for the task at hand. The attention method constructs a learnable mask for each attribute and learns each attribute separately based on the backbone network. However, this approach may result in unequal impacts of different masks on the main task. In contrast, our method weights the attention of different channels in the representation obtained by the main task, which essentially corrects the main task while avoiding potential inconsistency issues caused by the attention method.

We compare the performance of our method CHARMS with the attention method in all experiments and summarized the results in Table 7. The table shows that the attention method did not perform as well as our method on all datasets. Specifically, on the DVM dataset, which involves a complex downstream task of 129 classification tasks, the attention method constructed different attentions for different attributes, which confused the backbone network and led to a decrease in overall task performance.

This finding highlights the impracticality of using the attention mechanism alone to integrate the abundant information in tabular data into the image model. This further supports the effectiveness of our proposed approach.

B.2. Comparison with CLIP method

CLIP is pre-trained on a large amount of text and image pairs, which makes it able to map from text to images. Some previous studies have demonstrated that CLIP is able to transform tabular data to text for classification given the column names (Wang & Sun, 2022; Hegselmann et al., 2023). However, CLIP is heavily reliant on the semantic information contained within the text, so the semantics of attributes are inevitable.

Indeed, the setting of this paper is more general. We expect to transfer the tabular knowledge to the image modality during training to cope with the absence of expert knowledge during testing. Our method CHARMS aims to automatically extract the semantic information from the tabular and align it with the corresponding image channels without requiring explicit knowledge of the attribute's precise meaning. Specifically, as we show in Section 4.2, based on measuring the similarity across attributes and channels, OT discovers and aligns the attribute semantic automatically.

We conducted an experiment with CLIP. In this experiment, we converted the tabular data into text format, such as "length: 16". To ensure a fair comparison, we utilized CLIP from (Radford et al., 2021) with the ResNet50 backbone. The CLIP model consists of an image encoder and a textual encoder, and we connected a one-layer linear head for classification or regression after the image encoder. Two versions of CLIP were trained in our experiment. CLIP-LP means CLIP-LinearProb, which denotes the scenario where the two encoders are fixed, and only the classification head is trained. CLIP-FT means

Table 8: Comparison with CLIP method. Here CLIP-LP means two encoders are fixed, and only the classification head is trained. CLIP-FT means fine-tuning the entire CLIP network.

	DVM↑	SUN ↑	CelebA ↑	Adoption ↑	Pawpularity ↓	Avito ↓
CLIP-LP	0.7619	0.6918	0.7590	0.3047	20.1537	0.2972
CLIP-FT	0.8417	0.8333	0.8165	0.2935	42.9489	0.2940
CHARMS	**0.9175**	**0.8661**	**0.8220**	**0.3603**	**18.4314**	**0.2495**

CLIP-FineTune, on the other hand, involves fine-tuning the entire CLIP network. With the contrastive learning of the two modalities of the CLIP model, tabular knowledge is transferred to the image modality. By transforming the task into a language-to-vision knowledge transfer, the results were obtained in Table 8.

From the experiments, we can see that the performance of CLIP is not ideal. This is probably due to the fact that in tabular data, each column holds its own distinct meaning, and directly utilizing it as input to CLIP can lead to the loss of certain information. For instance, on the CelebA dataset, the attribute ”wood (not part of a tree)” might not be a highly significant feature. However, when this attribute is converted to text format, its character length tends to be relatively long, which can introduce redundancy in the information.

From another perspective, previous work has pointed out that there is a modality gap in the CLIP's embedding space (Liang et al., 2022). This gap is caused by a combination of model initialization and contrastive learning optimization. In a multi-modal model with two encoders, the representations of the two modalities are clearly apart when the model is initialized. During optimization, contrastive learning keeps the different modalities separate by a certain distance. This gap makes the CLIP method fail in our task.

In summary, the loss of information and the modality gap that arises when transferring tabular data to images can hinder the performance of the CLIP method in our setting. However, our method addresses these challenges by automatically discovering and establishing the matching relationship between the two modalities, thereby facilitating effective knowledge transfer, which is a more general method.

C. More Experiments

C.1. More Visualization

We provide more visualizations in Table 9 to validate the ability of CHARMS to match the corresponding attributes and channels. We apply GradCAM on various datasets, which show similar visualization results, where the channels could be matched to a certain attribute with semantic meaning.

For the Adoption dataset, all tabular attributes are inherently more abstract in nature. However, for the purpose of visualization, we have specifically selected features that are visually recognizable by humans from images. For instance, attributes such as the type of pet and the color of the pet highlight more general aspects that are of interest.

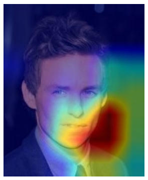
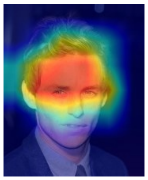
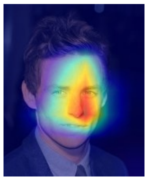
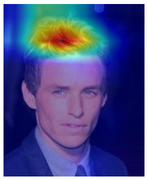
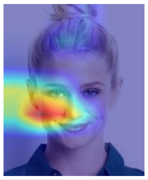
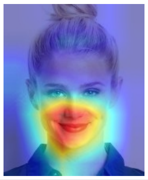
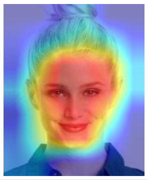
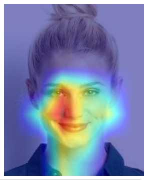
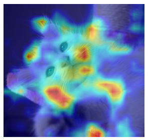
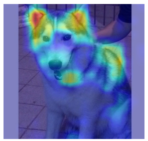
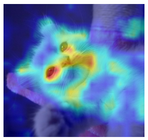
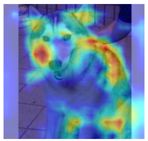
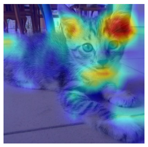
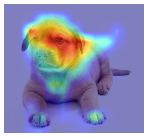
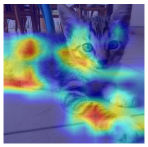
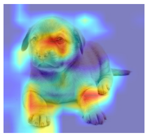
From the visualization, we can see that the judgment of the pet type focuses more on the pet's head, whereas the judgment of the color takes into account the whole body of the pet, and from this point of view we believe that our approach does achieve knowledge transfer.

C.2. Visualization with t-SNE

To visualize the impact of our method on the distribution of image features, we conducted experiments using the t-SNE method (Van der Maaten & Hinton, 2008). t-SNE can map high-dimensional data to a two- or three-dimensional space, enabling better visualization and interpretation of the data structure. The method employs a nonlinear mapping approach that minimizes the difference between the distances of points in high-dimensional space and those in low-dimensional space. Specifically, it represents high-dimensional data points as probability distributions and generates corresponding probability distributions in the low-dimensional space. Then, it uses KL divergence to measure the difference between the two probability distributions and minimizes it to achieve the best mapping effect.

The experimental results are presented in Figure 7, where the ORIGIN method refers to training with image modalities only.

Table 9: More Visualization by GradCAM.

Tabular Attribute	5_o_Clock_Shadow	Arched_Eyebrows	Big_Nose	Blond_Hair
Aligned Channel	65, 87, 119, 236...	33, 76, 78, 115, ...	50, 224, 258, ...	684
Visualization				
Tabular Attribute	High_Cheekbones	Smiling	Oval_Face	Rosy_Cheeks
Aligned Channel	2, 26, 41, 85,...	11, 12, 28, 57, ...	52, 646, 924, ...	4, 47, 88,...
Visualization				
Tabular Attribute	Type		Color	
Aligned Channel	399, 413, 414, 521...		400, 412, 425, 448...	
Visualization				
Aligned Channel	399, 413, 414, 521...		400, 412, 425, 448...	
Visualization				

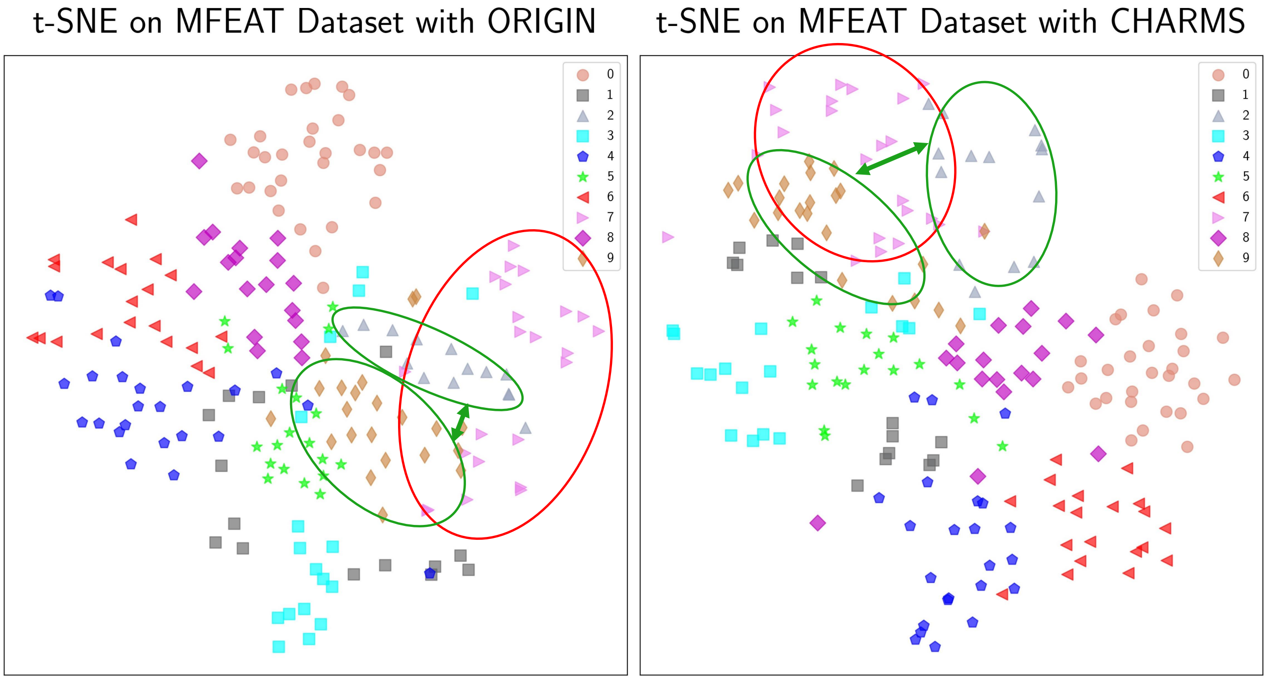


Figure 7: Visualization of t-SNE on the MFEAT dataset. the ORIGIN method represents training on image modalities only. As can be seen from the figure, our method makes the intra-class distance smaller and the inter-class distance larger. Therefore the transfer of expert knowledge from tabular data to the image model is effective. The red circles mean that our method makes the intra-class distance smaller, and the green circles indicate that our method makes the inter-class distance larger.

The figure shows that the ORIGIN method achieved good segmentation results due to the task's simplicity. However, due to the lack of expert knowledge, the intra-class distance is still large, particularly for samples with label 7, while the inter-class distances remain small, such as for samples with labels 2 and 9. In contrast, our method compensates for these deficiencies by transferring expert knowledge.

C.3. More Mutual Information experiments

We chose the MFEAT dataset for the Mutual Information experiments since, in this dataset, the formal features of each category are simple and easily distinguishable. For example, morphological features and non-morphological features. And the images are all digital images, which are relatively simple and easy to understand. The experiment mainly helps us understand. More mutual information experiments can be obtained in Figure 8 9.

The experiments in PetFinder-adoption dataset also indicate that existing methods for transferring tabular knowledge to image models yield low mutual information between the representations and tabular data. Our CHARMS method, on the other hand, maximises the mutual information of tabular and images to achieve better results.

C.4. More Ablation Studies

In the CHARMS method, we use the K-Means (Lloyd, 1982; MacQueen, 1967) method to cluster the 2048-dimensional features extracted from ResNet. We discuss the number of clusters on the SUNAttribute dataset, and the results in Table 10 show that the performance of CHARMS is not affected by the number of clusters taken, demonstrating the robustness of the method to hyperparameter choices. This robustness makes the method more flexible and reliable in practical applications, as it does not require excessive hyperparameter tuning or fine-tuning, saving time and effort.

To investigate the effectiveness of the OT method in our proposed approach, CHARMS, we conducted experiments where we reversed the transfer matrix of OT, expecting the image channels to learn the unaligned tabular attributes. We denote this approach as CHARMS-reverse. The results of this experiment are shown in Table 11, which demonstrate that the performance

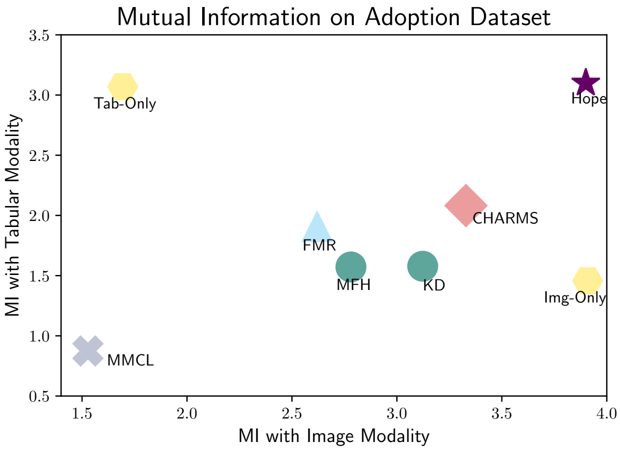


Figure 8: Mutual Information with Different Modality on the Adoption Dataset.

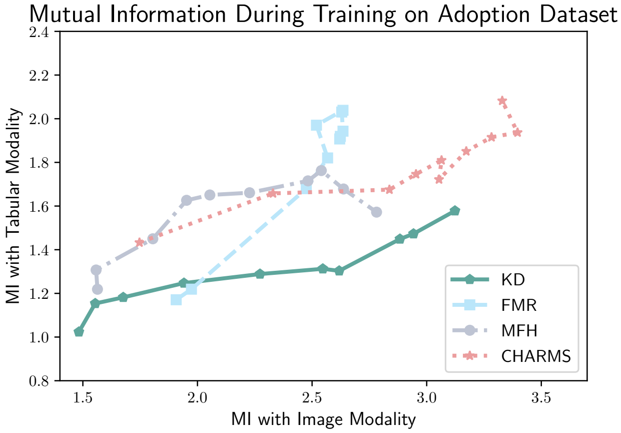


Figure 9: Mutual Information During Training on the Adoption dataset.

Table 10: Ablation study on cluster number on SUNAttribute dataset.

n_cluster	20	40	60	80	100
Accuracy	0.8494	0.8661	0.8494	0.8556	0.8522

of CHARMS-reverse is significantly lower than that of our original method, CHARMS, highlighting the importance of OT in alignment.

To further demonstrate the applicability and robustness of our proposed method, CHARMS, we conducted experiments using different network structures on DVM dataset with results shown in Table 12. The result also shows that the performance improvements achieved by our method are consistent across different network structures.

D. Limitations and Future Works

Our approach relies on leveraging mutual information between the two modalities, which establishes the feasibility of knowledge transfer. When there is a significant amount of mutual information present between the tabular and image modalities, our approach can effectively transfer relevant knowledge and insights between them. On the other hand, converting text into tables is indeed a viable approach, but this approach results in the loss of some of the textual information and it is challenging to handle such a conversion well. The problem of testing data drift also exists in real life. We will consider this problem deeply in future work. In terms of social impact, we think that our approach holds potential for application in the medical field, where it can assist doctors in making rapid and accurate diagnoses. There should be no negative social impact of our method.

Our work demonstrates the effectiveness of our method in both classification and regression tasks. In future work, it would be valuable to investigate the applicability of our method to other tasks, such as semantic segmentation. These types of tasks may require additional domain-specific knowledge, such as precise object localization within images, to achieve optimal

Table 11: Ablation study on Optimal Transport. CHARMS-reverse means that we reverse the transfer matrix of OT and make channels and attributes misaligned. The performance degradation proves that alignment is important.

	DVM ↑	SUN ↑	CelebA ↑	Adoption ↑	Pawpularity ↓	Avito ↓
CHARMS	0.9175	0.8661	0.8220	0.3603	18.4314	0.2495
CHARMS-reverse	0.8865	0.8459	0.8165	0.3440	18.8068	0.2568

Table 12: Impact of different network structures on the method on DVM dataset.

	ResNet	DenseNet	Inception	MobileNet
Model Size / M	25.8	8.2	6.8	3.7
ORIGIN	0.8743	0.8671	0.7492	0.8206
CHARMS	0.9175	0.9115	0.9012	0.8961

performance. Nonetheless, we believe that our approach is still applicable for such tasks.

On the other hand, the high cost of annotating expert data often leads to imbalanced datasets, which pose a challenge for improving image model performance using a limited amount of tabular data. Therefore, addressing this data imbalance is crucial for future work.