On Transferring Expert Knowledge from Tabular Data 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]
[1] National Key Laboratory for Novel Software Technology, Nanjing University, China
[2] School of Artificial Intelligence, Nanjing University, China
[3] Dept of Computer Science, School of EECS, Peking University
[4] School of Computer Science and Engineering, Nanjing University of Science and Technology
{jiangjp,yehj,jiangy,zhandc}@lamda.nju.edu.cn
leyewang@pku.edu.cn yyang@njust.edu.cn

Editors: Marco Fumero, Emanuele Rodolà, Clementine Domine, Francesco Locatello, Gintare Karolina Dziugaite, Mathilde Caron

Abstract

Transferring knowledge across modalities has garnered significant attention in the field of machine learning as it enables the utilization of expert knowledge from diverse domains. In particular, the representation of expert knowledge in tabular form, commonly found in fields such as medicine, can greatly enhance the comprehensiveness and accuracy of image-based learning. However, the transfer of knowledge from tabular to image data presents unique challenges due to the distinct characteristics of these data types, making it challenging to determine "how to reuse" and "which subset to reuse". To address this, we propose a novel method called CHannel tAbulaR alignment with optiMal tranSport (CHARMS) that automatically and effectively transfers relevant tabular knowledge. Specifically, by maximizing the mutual information between a group of channels and tabular features, our method modifies the visual embedding and captures the semantics of tabular knowledge. The alignment between channels and attributes helps select the subset of tabular data which contains knowledge to images. Experimental results demonstrate that CHARMS effectively reuses tabular knowledge to improve the performance and interpretability of visual classifiers.

1 Introduction

Data takes on various forms, such as images, text, video, and audio, providing rich and diverse sources of information for a given task. In contrast to using a single modality, multimodal learning aims to fuse information from different data modalities to create more comprehensive and accurate models [3; 41; 47; 64; 67; 62]. This approach has demonstrated exceptional performance across many domains, including recommender systems [48; 21; 2], healthcare [69; 14], and visual question answering [33; 70; 25].

In practical applications, obtaining data from multiple modalities can be challenging [71], as expert knowledge or specialized equipment may be required, such as medical images. The high acquisition cost of such data makes the traditional multimodal fusion approach impractical. To address this, one solution is to employ multiple modalities during training, enabling expert knowledge to transfer from one modality to another and improving the performance of a single modality during testing. The

Proceedings of the I edition of the Workshop on Unifying Representations in Neural Models (UniReps 2023).

current research on crossmodal transfer primarily focuses on images and text [26; 56; 46], but limited exploration has been done with tabular data [13].

Tabular data is a common type of structured data, usually organized in a table format, where each column represents an attribute or feature and each row represents a sample of data [40]. Tabular data often involves some expert knowledge, for example, in the medical field, an attribute of tabular data may represent position information in an MRI image that needs to be focused on, which requires expert annotation. Therefore, transferring expert knowledge from tables to images will improve detection efficiency and reduce the burden on doctors. However, tabular data's structured format distinguishes it from existing unstructured data such as text, making existing crossmodal transfer methods unsuitable for tabular data [28; 49].

Specifically, we face two challenges in transferring tabular knowledge for images. Firstly, we must address "how to reuse" the tabular data. As each column in tabular data has a unique semantic meaning, relying on standard RNN [18; 68] or Transformer [55] methods to construct a coarse feature space would result in a loss of interpretability of certain attributes. Moreover, categorical and numerical variables in tabular data require different processing methods. Secondly, we must identify "what subset to reuse" from the vast amount of information contained in tabular data since not all of it is 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. Therefore it is crucial to identify the useful information that can be transferred to instruct the learning of images. We expect that by transferring tabular knowledge to an image model, the model can learn corresponding semantics more effectively and achieve better performance on correlation tasks.

To overcome the aforementioned challenges, we propose a novel method named CHannel tAbulaR alignment with optiMal tranSport (CHARMS) that aligns tabular data attributes with image channels which automatically transfers relevant expert knowledge in tabular data to images. Specifically, we modify the visual embedding with the instruction of tabular data as auxiliary information and learning tabular features with a group of channels, maximizing the mutual information between them. Additionally, we utilize the optimal transport algorithm [6; 8; 66] to match the representation of each channel with the representation of each attribute, where a distinction is made between categorical and numerical variables. We strengthen the corresponding channels to ensure a focused learning of the tabular knowledge. In this way, our approach can automatically and effectively utilize expert knowledge from tabular data in the learning process, outperforming previous methods. To summarize, our contribution is three-fold:

- We emphasize the importance of knowledge transfer from tabular data to image data, as this can lead to improved performance when tabular data is missing due to high costs.
- We propose CHARMS method to automatically transfers 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 performance of visual classifiers. Moreover, our approach offers insightful explanations of the learned visual embedding space with tabular instruction.

This paper is organized as follows: the related work is introduced in Section 2. Section 3 and Section 4 provide the setting formalization, discovery experiment and our method. In Section 5, we present experiment results and discuss our findings. Finally, Section 6 concludes our study results.

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 of the task for different scenarios. An important task in multimodal learning is the fusion of modalities. Some previous work used BERT [29; 50] or co-attention [33; 51] to fuse different modal information simply. Subsequently, some large models [31; 24; 30] were created to align the information of different modalities in terms of their semantic relationships using contrastive learning approach [52]. Different pre-training approaches have also been extensively studied [4; 22; 65; 35].

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. For example, knowledge in missing modalities can be complemented using autoencoders or generative adversarial approaches [9; 42; 32]. Ma et al. [38] improves the robustness of Transformer models by automatically searching for an optimal fusion strategy regarding input data. Wang et al. [57] proposed a framework based on knowledge distillation, utilizing the supplementary information from all modalities, and avoiding imputation and noise associated with it. Hager et al. [13] 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. Traditional machine learning methods have been widely used on some tabular data, such as decision trees [45], support vector machines [54], and random forests [7]. 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 [58; 20; 12] 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 [11; 27; 44]. In recent years, more and more approaches for tabular data have been proposed [1; 17; 61; 23]. 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 some analysis.

3.1 Transfer Knowledge from Tabular to Images

Formally, we define the crossmodal transfer training dataset $D_{train} = \{\boldsymbol{x}_i^T, \boldsymbol{x}_i^I, y_i\}_{i=1}^N$, where $\boldsymbol{x}^I \in \mathbb{R}^{H_0 \times W_0 \times C_0}$ represent image data, $\boldsymbol{x}^T \in \mathbb{R}^D$ represent tabular data and $y \in Y$ is the label space of the task. The image data is represented as a three-dimensional tensor with height H_0, width W_0, and RGB channels $C_0 = 3$, while the tabular data is a vector of dimension D, where each dimension corresponds to an attribute. We define the test dataset $D_{test} = \{\boldsymbol{x}_i^I\}_{i=1}^M$, where tabular modality is missing due to high collection cost and the need for expert annotation. During training, we aim to minimize the empirical risk of model $f(\boldsymbol{x})$ over the training set:

$$\sum_{(\boldsymbol{x}_i^I, \boldsymbol{x}_i^T, y_i) \in D_{train}} \mathcal{L}(f(\boldsymbol{x}_i^I), y_i \mid \boldsymbol{x}_i^T), \tag{1}$$

where $\mathcal{L}$ is the loss function that measures the discrepancy between prediction and ground-truth label such as cross-entropy loss and $\mid$ indicates conditioning on the tabular data. The model can be decomposed into embedding and linear classifier: $f(\boldsymbol{x}) = \boldsymbol{W}^\top \phi(\boldsymbol{x})$, where $\phi(\cdot) : \mathbb{R}^D \to \mathbb{R}^d$ is the feature extractor to extract the embedding of the images and $\boldsymbol{W} \in \mathbb{R}^{d \times |Y|}$.

Our objective is to transfer relevant tabular information into the image model f. In situations where expert knowledge is not available, we expect the model to provide better predictions when only given the image data $\boldsymbol{x}^I$ on the test set.

3.2 Methods for Crossmodal Transfer

One of the main challenges in this task is how to transfer the tabular knowledge to the image model. It is feasible to align the two modality and then select the appropriate part for knowledge transfer. So we explore methods with alignment from different perspectives, including output-based transfer, parameter-based transfer, and embedding-based transfer.

Output-based Transfer. To transfer knowledge from tabular data to image models, we aim to ensure that the predictions of image model f and tabular model g are aligned. To achieve this, we first train a classifier g on the tabular data such as LightGBM [27]. We then fit the prediction results of the image model f to g during the training. Knowledge Distillation (KD) [16] is an output-based method:

$$\mathcal{L}(\boldsymbol{x}^I, \boldsymbol{x}^T, y) = (1-\lambda)\mathcal{L}(\boldsymbol{x}^I, y) + \lambda\mathcal{L}_{\mathrm{KD}}(f(\boldsymbol{x}^I), g(\boldsymbol{x}^T)). \tag{2}$$

$\mathcal{L}_{\mathrm{KD}}$ measures the similarity between the prediction of two models with Kullback-Leibler (KL) divergence g is called teacher network and f is student network. Aligning the output of the tabular model and the current model helps to reuse the knowledge in tabular data.

So as Modality Focus Hypothesis (MFH) [60], the modality general decisive information is set according to the feature importance [7; 59] in tabular data as the teacher network, selecting subset of the tabular data. Then only use $\mathcal{L}_{\mathrm{KD}}$ for distillation to fully observe the tabular's influence on image.

Parameter-based Transfer. The parameters of the model may contain part of the knowledge in the data, so the knowledge can be transferred from the perspective of the parameters of the model as well. For example, Fixed Model Reuse (FMR) [63] utilizes the learning power of deep models to implicitly grab the useful discriminative information from fixed models/features. In our setting, the fixed features referred to here are the tabular data:

$$\mathcal{L} = y \log h\left(f\left(\boldsymbol{x}^I\right) + g\left(\boldsymbol{x}^T\right)\right) + \frac{1}{2}\left\|\boldsymbol{x}^T - \phi(\boldsymbol{x}^I)\boldsymbol{U}\right\|_F^2. \tag{3}$$

h is a soft-max operator and $\boldsymbol{U}$ is the linear connections between the tabular features and embedding of images. To transfer the influence of the fixed features $\boldsymbol{x}^T$ to images during the training procedure, FMR removes those connected parts corresponding to features $\boldsymbol{x}^T$ gradually and finally vanish all related components with the knockdown method.

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) [13] proposes the self-supervised contrastive learning framework that takes advantage of images and tabular data to train unimodal encoders:

$$\begin{aligned} &\mathcal{L} = \lambda\ell_{I,T} + (1-\lambda)\ell_{T,I}, \quad z_{j_I} = f_{\phi_I}\left(\phi(\boldsymbol{x}^I)\right), \\ &\ell_{I,T} = -\sum_{j\in\mathcal{N}} \log \frac{\exp\left(\cos\left(z_{j_I}, z_{j_T}\right)/\tau\right)}{\sum_{k\in\mathcal{N}, k\neq j}\exp\left(\cos\left(z_{j_I}, z_{k_T}\right)/\tau\right)}, \end{aligned} \tag{4}$$

where embeddings are propagated through separate projection heads f_{ϕ_I} and f_{ϕ_T} and brought into a shared latent space as projections x_{j_I}, z_{j_T}. $\ell_{I,T}$ is calculated analagously. $\mathcal{N}$ denotes all subjects in a batch. Then MMCL uses linear probing of frozen networks to evaluate the quality of the learned representations. By mapping tabular and image data to the same space and utilizing contrastive learning methods, the knowledge in tabular data can be transferred into an image feature extractor.

While the output-based, parameter-based, and embedding-based methods offer perspectives on transferring knowledge between modalities, each method has its own limitations. The output-based approach offers a simple and straightforward alignment based on the output of the model, but it may not capture detailed information for a certain attribute. The MFH method considers important features, but it completely discards other information during knowledge distillation. Parameter-based methods such as FMR cannot address the significant differences between tabular and image models, and the information contained in the parameters may be limited. 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. By exploring these different transfer methods and their respective limitations, we can gain a deeper understanding of the challenges and opportunities in multimodal learning and develop more effective approaches for transferring knowledge from table to images.

4 Transferring Knowledge after Alignment

Motivated by the unique characteristics of tabular data, we leverage it as auxiliary information in our approach to transfer knowledge to the image modality. Specifically, we minimize the mutual information between the image and each attribute of the table data, effectively transferring the relevant

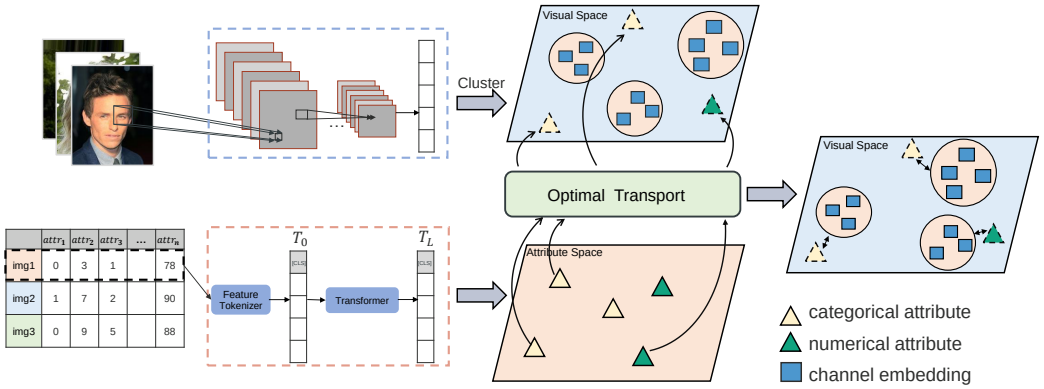

Figure 1: Flow chart of CHARMS method. Our approach combines the learning of image and tabular data, leveraging the specific characteristics of each modality to effectively transfer knowledge from one to the other. We use Optimal Transport (OT) methods to match tabular attributes to image channels, effectively 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.

table knowledge to the image modality. Additionally, we use Optimal Transport to match the expert knowledge that can be expressed in the image data, allowing us to select a subset of the image features and strengthen the learning of the corresponding channels. Our approach highlights the importance of leveraging the specific characteristics of each modality to develop effective transfer. The flowchart is shown in Figure 1.

4.1 Channel Tabular Alignment

To extract the relevant information from the tabular data that is beneficial to the image model, we also use alignment-based methods for feature selection. This task consists of two main parts: first, obtaining the intermediate embedding of the image and tabular data; and second, performing alignment-based feature selection.

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 $\boldsymbol{x}^{\boldsymbol{I}}$ are defined as $\phi_{-1}(\boldsymbol{x}^{\boldsymbol{I}}) \in \mathbb{R}^{H\times W\times C}$, where C is the number of channels, and each channel corresponds to a high-level feature such as edges, whose shape is $H \times W$.

Similarly, 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 $\boldsymbol{x}^T_{\text{cat}}$.

Secondly, we use the optimal transport to align the channels of the image with the attributes of the tabular data [5]. 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. The basic idea behind OT is to find a mapping between the elements of two distributions that minimizes the cost of moving one distribution to the other. The cost is typically defined as a distance metric between the elements. However, not all tabular attributes can be displayed on the image, and in some cases, there may be missing or irrelevant attributes that cannot be aligned with the image data. For example, on the PetFinder-adoption dataset, the photo of the pet can reflect the pet's hair, body size, and other attributes, but not the health condition or vaccination status. To address this issue, we use the partial optimal transport (POT) algorithm [10].

Specifically, To address the issue that different channels of an image may have repeated semantics with some redundancy, we use K-Means [37; 39] clustering to group similar channels together. This allows us to obtain fewer distinct channels, each capturing a distinct aspect of the image data. Then we compute the cosine similarity of the dataset on each channel, resulting in a matrix $\boldsymbol{S_I} \in \mathbb{R}^{C'\times N\times N}$,

where C' is the number of clustered channels and N is the length of the dataset. In parallel, we process the attributes of the tabular data similarly to obtain the attribute-wise similarity matrix $\boldsymbol{S_T} \in \mathbb{R}^{D\times N\times N}$. Then the cost matrix is constructed from the channel-wise similarity between attribute-wise similarity. Then the OT transfer matrix is calculated:

$$\boldsymbol{C_{ij}} = \left\|\boldsymbol{S_{T_i}} - \boldsymbol{S_{I_j}}\right\|_2^2, \quad \boldsymbol{T} = \underset{\boldsymbol{T}}{\arg\min}\langle \boldsymbol{C}, \boldsymbol{T}\rangle_F, \tag{5}$$

where $\langle\cdot\rangle_F$ denotes the Frobenius norm. After aligning the distributions of the image and tabular data, we obtain the transfer matrix $\boldsymbol{T} \in \mathbb{R}^{D\times C'}$. Based on the clustering results, we can restore the corresponding relationship between the tabular attributes and the original channels of the image as $\boldsymbol{A} \in \mathbb{R}^{D\times C}$. Then the channels and attributes are aligned and relevant features are selected.

4.2 Learning with Auxiliary Information

To leverage the knowledge of each attribute of the tabular data, we construct auxiliary tasks to learn this information. Specifically, we use the matrix $\boldsymbol{A}$ to weigh the image channels, allowing us to focus the attention of the relevant tabular attributes on the corresponding image channels. We use the feature extractor of an existing image network $\phi(\cdot)$ to learn a classifier that maps from an attention image to the corresponding attributes of the tabular data. By doing so, we enhance the image network's understanding of the attributes of the tabular data and transfer this knowledge into the image modality. This allows the learned model to handle missing tabular modalities and improve its overall performance on complex tasks.

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

$$\begin{aligned}
&\mathcal{L} = \mathcal{L}(f(\boldsymbol{x}^I), y) + \mathcal{L}(g(\boldsymbol{x}^T), y) + \mathcal{L}_{i2t}, \\
&\mathcal{L}_{i2t} = \sum_p \mathcal{L}_{MSE}(\mathcal{A}_p \cdot \phi(\boldsymbol{x}^I), \boldsymbol{x}^T_{\text{num}p}) + \sum_q \mathcal{L}_{CE}(\mathcal{A}_q \cdot \phi(\boldsymbol{x}^I), \boldsymbol{x}^T_{\text{cat}q}).
\end{aligned} \tag{6}$$

Here, $\mathcal{L}$ is the label prediction loss function such as cross entropy loss for classification tasks. Since there may be numerical and categorical attributes for tabular data, we model them separately when constructing the loss to guide the image model to learn more information, expecting that the processing of different types is reasonable. $\mathcal{L}_{CE}$ is cross entropy loss for categorical attributes and $\mathcal{L}_{MSE}$ is mean square error loss for numerical attributes. This style of updating ensures that the model learns increasingly accurate channel-attribute correspondences, allowing the tabular data to guide the image data with increasing precision. By leveraging this approach, we can effectively transfer expert knowledge to images to develop more accurate and comprehensive image models for complex tasks.

To sum up, our method leverages OT to align the distributions of different modalities and select relevant tabular attributes that are closely related to the image data. We then use the alignment to enhance the image learning of the relevant attributes, thus transferring expert knowledge from the tabular data to the image model.

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 result of the alignment of attributes and channels.

5.1 Experiments and Results

Dataset. Totally six datasets are used in the experiment: **Data Visual Marketing (DVM)** [19] 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 [13], only the new version DVM dataset is available. Car models with less than 700 samples were removed, resulting in 129 target classes, a classification task. **SUNAttribute** [43]: 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** [36] is the abbreviation of CelebFaces Attribute, meaning

Table 1: Comparisons with baseline methods on DVM, SUN, CelebA and Adoption datasets. All tasks are classification tasks. RTDL means the FT-transformer [12] model trained on the tabular modality.

	DVM ↑	SUN ↑	CelebA ↑	Adoption ↑
LGB	0.9748	0.8501	0.7963	0.4101
RTDL	0.9682	0.8563	0.7936	0.4107
Resnet	0.8743	0.8361	0.8146	0.3477
KD	0.8390	0.8382	0.8118	0.3532
MFH	–	0.8312	0.7507	0.3041
FMR	0.8427	0.8347	0.8003	0.3526
MMCL	0.8203	0.8431	0.8041	0.2981
CHARMS	**0.9175**	**0.8661**	**0.8220**	**0.3603**

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.

Evaluation metrics. For all our classification tasks, we use accuracy to measure the performance.

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 supplementary material.

Results. To demonstrate the superiority of CHARMS, we compare it with other popular methods on six datasets as shown in Table 1. Our results show that CHARMS consistently achieves the best performance on all datasets. In contrast, the baseline methods we compared with do not significantly improve the performance compared to direct training with images. In fact, some of them even decrease the results. This is likely because these methods only use the tabular data to guide the image model at a coarse level, without considering the complex relationships and interactions between the modalities. As a result, the guidance provided by these methods is not sufficient for the image model to learn useful information, which can lead to confusion and poor results.

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 DVM 129 classification task.

What is particularly surprising about our approach is that it can outperform the tabular modality on the SUNAttribute dataset. Similarly, on the CelebA 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 visualize the impact of our method on the distribution of image features, we conducted experiments using the t-SNE method [53]. 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 2, where the ORIGIN method refers to training with image modalities only. 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

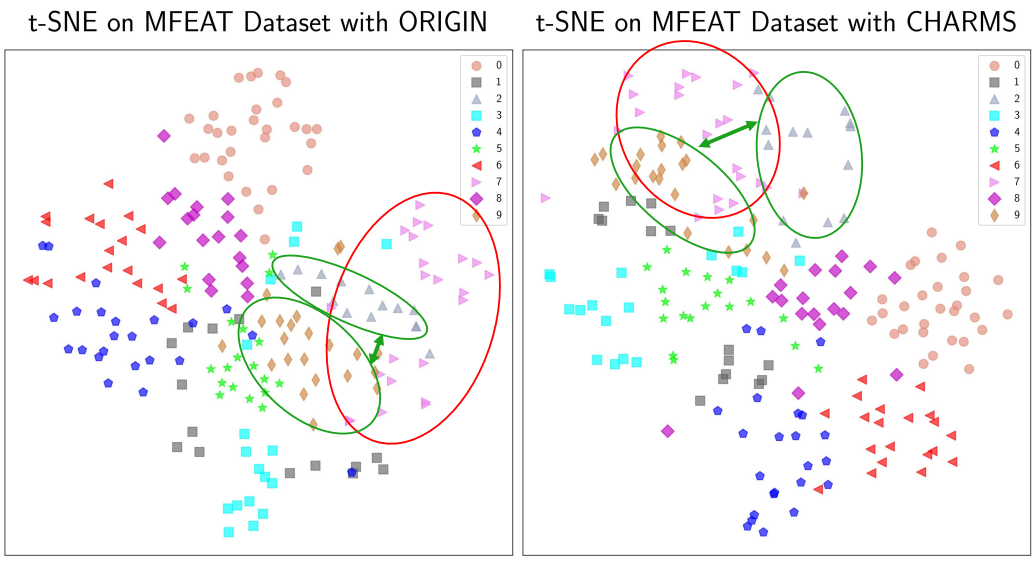

Figure 2: Visualization of t-SNE on the MFEAT dataset. the ORIGIN method represents training on image modalities only. As can be seen from the figure, our method makes the intra-class distance smaller and the inter-class distance larger. Therefore the transfer of expert knowledge from tabular data to 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.

as for samples with labels 2 and 9. In contrast, our method compensates for these deficiencies by transferring expert knowledge.

5.2 Experiments Analysis

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:

$$\phi(\boldsymbol{x}^T)_{att} = \mathcal{T}(\phi(\boldsymbol{x}^T)) \cdot \phi(\boldsymbol{x}^T) \tag{7}$$

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 2. The table shows that the attention method did not perform as well as our method. 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.

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 [58; 15]. However,

Table 2: 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 ↑
Attention	0.4757	0.8550	0.8180	0.3454
CHARMS	**0.9175**	**0.8661**	**0.8220**	**0.3603**

Table 3: 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 ↑
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**

CLIP is heavily reliant on the semantic information contained within the text, so that the semantics of attributes are inevitable.

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 [46] 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 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 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 3.

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 [34]. 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 different modalities separate by a certain distance. This gap makes CLIP 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. 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.

6 Conclusion

In this work, we propose the CHARMS, a novel 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.

References

[1] Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687, 2021.

[2] Paul Baltescu, Haoyu Chen, Nikil Pancha, Andrew Zhai, Jure Leskovec, and Charles Rosenberg. Itemsage: Learning product embeddings for shopping recommendations at pinterest. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2703–2711, 2022.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41:423–443, 2018.

[4] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

[5] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37:A1111–A1138, 2015.

[6] Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pages 1–12, 2011.

[7] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[8] Luis A Caffarelli and Robert J McCann. Free boundaries in optimal transport and monge-ampere obstacle problems. *Annals of mathematics*, 171:673–730, 2010.

[9] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1158–1166, 2018.

[10] Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33:2903–2913, 2020.

[11] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[12] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

[13] Paul Hager, Martin J Menten, and Daniel Rueckert. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. *arXiv preprint arXiv:2303.14080*, 2023.

[14] Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20707–20717, 2022.

[15] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581, 2023.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[17] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

[18] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79:2554–2558, 1982.

[19] Jingmin Huang, Bowei Chen, Lan Luo, Shigang Yue, and Iadh Ounis. Dvm-car: A large-scale automotive dataset for visual marketing research and applications. In *2022 IEEE International Conference on Big Data*, pages 4140–4147, 2022.

[20] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

[21] Zhenhua Huang, Xin Xu, Juan Ni, Honghao Zhu, and Cheng Wang. Multimodal representation learning for recommendation in internet of things. *IEEE Internet of Things Journal*, 6:10675–10685, 2019.

[22] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 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, pages 12976–12985, 2021.

[23] Alan Jeffares, Tennison Liu, Jonathan Crabbé, Fergus Imrie, and Mihaela van der Schaar. Tangos: Regularizing tabular neural networks through gradient orthogonalization and specialization. *arXiv preprint arXiv:2303.05506*, 2023.

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916, 2021.

[25] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11181–11188, 2020.

[26] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[27] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[28] Ralph Kimball and Margy Ross. *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.

[29] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11336–11344, 2020.

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900, 2022.

[31] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[32] Linchao Li, Bowen Du, Yonggang Wang, Lingqiao Qin, and Huachun Tan. Estimation of missing values in heterogeneous traffic data: Application of multimodal deep learning model. *Knowledge-Based Systems*, 194:105592, 2020.

[33] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[34] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[35] Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, 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*, pages 6008–6018, 2020.

[36] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, pages 3730–3738, 2015.

[37] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28:129–137, 1982.

[38] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021.

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

[40] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, pages 51–56, 2010.

[41] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning*, pages 689–696, 2011.

[42] Yongsheng Pan, Mingxia Liu, Yong Xia, and Dinggang Shen. 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.

[43] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108:59–81,

2014.
[44] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.
[45] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.
[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021.
[47] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34:96–108, 2017.
[48] Aghiles Salah, Quoc-Tuan Truong, and Hady W Lauw. Cornac: A comparative framework for multimodal recommender systems. *The Journal of Machine Learning Research*, 21:3803–3807, 2020.
[49] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.
[50] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.
[51] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
[52] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.
[53] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
[54] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10:988–999, 1999.
[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
[56] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.
[57] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1828–1838, 2020.
[58] Zifeng Wang and Jimeng Sun. Transtab: Learning transferable tabular transformers across tables. *arXiv preprint arXiv:2205.09328*, 2022.
[59] Maksymilian Wojtas and Ke Chen. Feature importance ranking for deep learning. *Advances in Neural Information Processing Systems*, 33:5105–5114, 2020.
[60] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*, 2022.
[61] Jiahuan Yan, Jintai Chen, Yixuan Wu, Danny Z Chen, and Jian Wu. T2g-former: organizing tabular features into relation graphs promotes heterogeneous feature interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10720–10728, 2023.
[62] Yang Yang, Han-Jia Ye, De-Chuan Zhan, and Yuan Jiang. Auxiliary information regularized machine for multiple modality feature learning. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 1033–1039, 2015.
[63] Yang Yang, De-Chuan Zhan, Ying Fan, Yuan Jiang, and Zhi-Hua Zhou. Deep learning for fixed model reuse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2831–2837, 2017.
[64] Yang Yang, De-Chuan Zhan, Yuan Jiang, and Hui Xiong. Reliable multi-modal learning: A survey. *Journal of Software*, 32:1067–1081, 2020.
[65] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021.
[66] Han-Jia Ye, De-Chuan Zhan, Yuan Jiang, and Zhi-Hua Zhou. Rectify heterogeneous models with semantic mapping. In *International Conference on Machine Learning*, pages 5630–5639. PMLR, 2018.

[67] Han-Jia Ye, De-Chuan Zhan, Xiaolin Li, Zhen-Chuan Huang, and Yuan Jiang. College student scholarships and subsidies granting: A multi-modal multi-label approach. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 559–568. IEEE, 2016.

[68] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

[69] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2418–2428, 2022.

[70] Chen Zheng, Quan Guo, and Parisa Kordjamshidi. Cross-modality relevance for reasoning on language and vision. *arXiv preprint arXiv:2005.06035*, 2020.

[71] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5:44–53, 2018.