AFFAKT: A Hierarchical Optimal Transport Based Method for Affective Facial Knowledge Transfer in Video Deception Detection

Zihan Ji[1], **Xuetao Tian**[2], **Ye Liu**[1*]

[1]School of Future Technology, South China University of Technology, Guangzhou, China
[2]Faculty of Psychology, Beijing Normal University, Beijing, China
ftjizihan@mail.scut.edu.cn, xttian@bnu.edu.cn, yliu03@scut.edu.cn

Abstract

The scarcity of high-quality large-scale labeled datasets poses a huge challenge for employing deep learning models in video deception detection. To address this issue, inspired by the psychological theory on the relation between deception and expressions, we propose a novel method called AFFAKT in this paper, which enhances the classification performance by transferring useful and correlated knowledge from a large facial expression dataset. Two key challenges in knowledge transfer arise: 1) how much knowledge of facial expression data should be transferred and 2) how to effectively leverage transferred knowledge for the deception classification model during inference. Specifically, the optimal relation mapping between facial expression classes and deception samples is firstly quantified using proposed H-OTKT module and then transfers knowledge from the facial expression dataset to deception samples. Moreover, a correlation prototype within another proposed module SRKB is well designed to retain the invariant correlations between facial expression classes and deception classes through momentum updating. During inference, the transferred knowledge is fine-tuned with the correlation prototype using a sample-specific re-weighting strategy. Experimental results on two deception detection datasets demonstrate the superior performance of our proposed method. The interpretability study reveals high associations between deception and negative affections, which coincides with the theory in psychology.

Introduction

Video deception detection has attracted a huge interest in various fields including law enforcement, jurisprudence, national security, business and interviewing (Chebbi and Jebara 2023). In the earlier days, researchers proposed several statistic-based methods (Jaiswal, Tabibu, and Bajpai 2016; Rill-García et al. 2019; Mathur and Matarić 2020), which utilize facial features, such as OpenFace (Baltrusaitis et al. 2018), action units (Jaiswal, Tabibu, and Bajpai 2016) and face landmarks (Rill-García et al. 2019), to achieve classification by applying machine learning methods on facial features with statistically significant differences. The performance of these methods heavily relies on the expert knowledge to construct and select valid feature sets. Hereafter, several deep learning based methods (Zhang et al. 2022; Hsiao and Sun 2022; Guo et al. 2023) employed different advanced neural networks to automatically extract powerful feature representations for video deception detection. However, their performances highly depends on the availability of high-quality labeled real data (Chen et al. 2019). Moreover, current datasets, *e.g.*, Real Life Trial (RTL) (Pérez-Rosas et al. 2015), DOLOS (Guo et al. 2023), typically contain small number of annotated samples, which poses limitations on training deep neural networks, thereby hindering further performance improvement. Taking PECL(visual) on DOLOS (Guo et al. 2023) as an example (Fig. 1, refer to appendix for detailed settings), PECL(visual) cannot obtain enough knowledge about detecting deception due to limited data and poor feature representation. Consequently, a key question in deception detection is how to develop a superior deep learning based method when only limited labeled deception data is available.

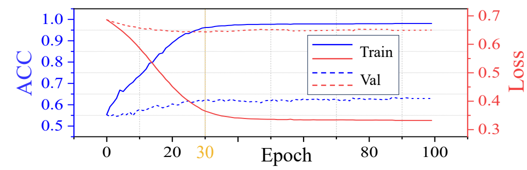


Figure 1: ACC and loss remain unchanged after 30 epochs.

To address this issue, motivated by the insights from psychological theory and previous researches that certain facial movements and expressions are associated with deception (Le 2016; Zuckerman, DePaulo, and Rosenthal 1981; DePaulo et al. 2003; Zloteanu 2020), we propose a hierarchical optimal transport (Guo et al. 2022) based method for AFfective FAcial Knowledge Transfer (AFFAKT). The objective of AFFAKT is to transfer and leverage the knowledge properly from a large-scale video facial expression recognition (VFER) dataset (*source domain*) to enhance video deception detection (*target domain*) model, enabling better detection performance. Subsequently, there are two key challenges when transferring knowledge: 1) *how much* knowledge of facial expression data should be transferred and 2) *how to* effectively leverage transferred knowledge for the deception classification model during inference.

Two modules are integrated in AFFAKT to handle these two challenges. Firstly, Hierarchical Optimal Transport Knowledge Transfer (H-OTKT) module is devised to auto-

*The corresponding author.

matically quantify the potential correlation between the class of facial expressions and deception samples via hierarchical optimal transport (H-OT). Such correlation mapping would be employed to determine how much knowledge from different categories is transferred to each sample. The transferred knowledge is further integrated with extracted target deception feature to enhance its representation. Secondly, Sample-specific Re-weighting Knowledge Bank (SRKB) module is proposed to learn the invariant class-level relation between VFER and video deception detection datasets via momentum updated correlation prototype whilst training. During testing phase, for each test sample, its correlation mapping with facial expression classes calculated by H-OTKT is fine-tuned using the correlation prototype via a sample-specific re-weighting strategy, so that inaccurate correlation can be alleviated. This strategy facilitates the robustness of the estimated relation mapping, and takes advantage of the learned correlated knowledge for detection inference more efficiently. The experimental results on two video deception detection datasets demonstrate the superior performance of our proposed method. And the interpretability studies reveal high associations between deception and negative affections, which coincide with the theory in psychology. The appendix can be found at (Ji, Tian, and Liu 2024). The code can be found at `https://github.com/Zander-J/AFFAKT`.

The main contributions of this paper can be summarized as follows:

- We propose a new method called AFFAKT, which facilitates the classification performance in video deception detection by transferring and leveraging the correlated knowledge from a large-scale facial expression dataset.
- An H-OTKT module is presented to quantify the correlation mapping between facial expression classes and deception samples, so as to transfer appropriate amount of knowledge from different source classes to deception samples. Besides, an SRKB module is exploited to enhance the correlation mapping with correlation prototype through a sample-specific re-weighting strategy during inference. SRKB effectively leverages transferred knowledge, thereby improving deception detection accuracy.
- Experiments including comparison experiments, ablation studies and interpretability analysis are performed on two video deception detection datasets. The superior performance validate the effectiveness of our proposed strategies. The interpretability studies reveal high associations between deception behavior and negative affections.

Related Work

Video Deception Detection. Video deception detection is one of the main tasks in affective computing and psychological researches (Krishnamurthy et al. 2018; Pérez-Rosas et al. 2015; Borza, Itu, and Danescu 2018). Earlier methods relied on manually selected statistical features including OpenFace (Jaiswal, Tabibu, and Bajpai 2016; Rill-García et al. 2019), action units (Jaiswal, Tabibu, and Bajpai 2016; Avola et al. 2019), or gestures (Şen et al. 2020) to achieve deception detection. Recently, the powerful representation capability of end-to-end deep learning has greatly improved the accuracy of deception detection. (Karnati et al. 2022; Ding et al. 2019; Krishnamurthy et al. 2018; Guo et al. 2023) used ResNet to encode video frames and LSTM to capture temporal information. (Guo et al. 2023) applied transformer (Vaswani et al. 2017) based ViT (Dosovitskiy et al. 2020) with adapter tuning (Houlsby et al. 2019) as temporal encoder to perform deception detection, leading to significant improvement of classification performance. However, one of the key problems is that the high-quality labeled dataset is always limited due to the initiative of lying and the expensive cost of annotation (Snchez-Junquera et al. 2020), which seriously hinders the development of deep learning methods.

Optimal Transport. Optimal transport (OT) (Peyré and Cuturi 2019) is a mathematical framework, seeking the most efficient way of transporting one distribution of mass into another. Let $p = \sum_{i=1}^{n} a_i \delta_{\mathbf{X}_{A_i}}$ and $q = \sum_{j=1}^{m} b_j \delta_{\mathbf{X}_{B_j}}$ be n and m dimensional discrete probability distributions for two finite sets $\mathbf{X}_A = \{\mathbf{X}_{A_i}\}_{i=1}^{n}, \mathbf{X}_B = \{\mathbf{X}_{B_j}\}_{j=1}^{m}$ respectively, where $\boldsymbol{a} \in \Delta_n$ and $\boldsymbol{b} \in \Delta_m$, Δ_n and Δ_m are the probability simplex of $\mathbb{R}^n$ and $\mathbb{R}^m$, and $\delta_{\mathbf{X}_*}$ refers to a point mass located at coordinate $\mathbf{X}_* \in \mathbb{R}^d$. Denoting $\mathbf{M} \in \mathbb{R}_+^{n\times m}$ as the cost matrix with $\mathbf{M}_{i,j} = \mathcal{M}(\mathbf{X}_{A_i}, \mathbf{X}_{B_j})$, which means the cost to transport one unit of mass between elements of the sets. Then, the transport plan matrix $\mathbf{T}$ is obtained by solving:

$$\mathrm{OT}(p,q) = \min_{\mathbf{T}\in\Pi(p,q)} < \mathbf{T}, \mathbf{M} >_{\mathrm{F}} \tag{1}$$

where $< \cdot, \cdot >_{\mathrm{F}}$ is the Frobenius dot-product. The constrain $\Pi(p,q) := \{\mathbf{T} \in \mathbb{R}_+^{n\times m} | \sum_i^n \mathbf{T}_{i,j} = b_j, \sum_j^m \mathbf{T}_{i,j} = a_i\}$ enforces $\mathbf{T}$ to have p, q as its marginals. It should be noted that $\mathbf{T}$ can be interpreted as the probabilistic correspondence between the elements of p and q. If the transport cost $\mathbf{M}_{i,j}$ between $\mathbf{X}_{A_i}$ and $\mathbf{X}_{B_j}$ is high, then a low correlation $\mathbf{T}_{ij}$ should be obtained. Eq. (1) is a linear assignment problem, which is expensive to solve. Fortunately, a entropy regularized OT has been developed as follows:

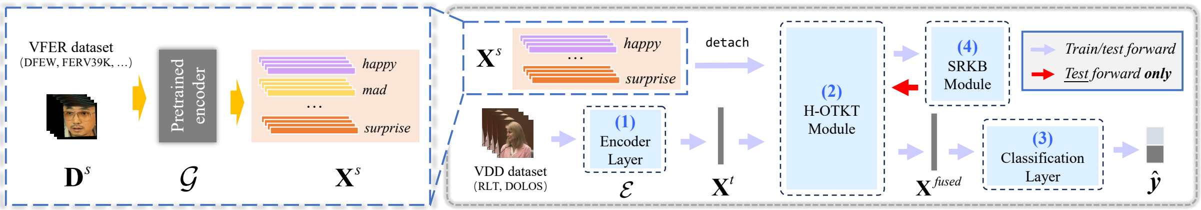
$$\mathrm{OT}(p,q) = \min_{\mathbf{T}\in\Pi(p,q)} < \mathbf{T}, \mathbf{M} >_{\mathrm{F}} - \epsilon\mathcal{H}(\mathbf{T}) \tag{2}$$

Here, $\mathcal{H}(\mathbf{T}) = -\mathbf{T}\log\mathbf{T}$ is the entropic regularization. Eq. (2) can be solved by the Sinkhorn algorithm efficiently (Cuturi 2013).

Hierarchical Optimal Transport. Hierarchical optimal transport usually contains high-level and low-level OT, where high-level OT learn the optimal transport plan with a given cost matrix, and the given cost matrix depends on the solution of low-level OT. Recently, H-OT has been recently studied for various tasks including multimodal distribution alignment (Lee et al. 2019), few-shot learning (Guo et al. 2022). For example, in (Lee et al. 2019), H-OT was used to leverage cluster structure in data to improve alignment in noisy, ambiguous, or multimodal settings. (Guo et al. 2022) proposed a novel distribution calibration method for few-show learning, where an adaptive weight matrix representing the relations between the base classes and novel samples is computed by hierarchical optimal transport.

The Proposed Method

Our proposed method AFFAKT for video deception detection contains four modules shown in Fig. 2 (b): (1) Encoder



(a) Extract source features (b) The pipeline of our proposed method AFFAKT

Figure 2: (a) Source features are extracted by a pre-trained encoder in advance. (b) The pipeline of our proposed AFFAKT. Four modules in AFFAKT are in blue background.

layer, (2) Hierarchical Optimal Transport Knowledge Tranfer (H-OTKT) module, (3) Classification layer, and (4) Sample-specific Re-weighting Knowledge Bank (SRKB) module. Assuming a video deception detection dataset is denoted as $\mathbf{D}^t = \{(\mathbf{V}_i^t, y_i^t) | \mathbf{V}_i^t \in \mathbb{R}^{F\times 3\times H\times W}, y_i^t \in \{1,\ldots,L^t\}\}_{i=1}^{N^t}$, where $(\mathbf{V}_i^t, y_i^t)$ is the i-th video sample and its ground truth label, N^t represents the number of samples in $\mathbf{D}^t$, F is the number of video frames, H and W are the height and width of video frames, L^t is the number of target categories (*i.e.,* deceptive and truthful). Our idea is to improve the classification performance on $\mathbf{D}^t$ by transferring useful and correlated knowledge from a large-scale VFER dataset $\mathbf{D}^s = \{(\mathbf{V}_j^s, y_j^s)\}_{j=1}^{N^s}$, where $y_j^s \in \{1,\ldots,L^s\}$, L^s is the number of categories in $\mathbf{D}^s$, and N^s is the number of samples in $\mathbf{D}^s$. Each category stands for one expression. In order to utilize $\mathbf{D}^s$ efficiently, a pre-trained encoder $\mathcal{G}$ is firstly employed to extract VFER feature representation $\mathbf{X}^s \in \mathbb{R}^{N^s\times d}$ in advance, *i.e.*, $\mathbf{X}^s = \mathcal{G}(\mathbf{V}^s)$, where d is the embedding dimension. By grouping $\mathbf{X}^s$ with the ground truth labels, $\mathbf{X}^s = \{\mathbf{X}^{s,k} \in \mathbb{R}^{J_k\times d}\}_{k=1}^{L^s}$ with $\sum_{k=1}^{L^s} J_k = N^s$, where $\mathbf{X}^{s,k}$ stands for feature embeddings of J_k samples in the k-th class. This process is shown in Fig. 2 (a). Pseudo-code of AFFAKT, all the symbol notations and their descriptions used in this paper are summarized in appendix.

Encoder Layer

For a video deception detection dataset, widely-used video and audio pre-trained models $\mathcal{E}$, *i.e.*, VideoMAE (Tong et al. 2022) and W2V2 (Baevski et al. 2020; Guo et al. 2023), are employed to learn visual and audio features separately. In our settings, there are 12 layers and 4 layers in VideoMAE and W2V2 respectively, where each layer contains a multi-head self-attention module and a feed-forward network. Since the number of samples in deception dataset is too small, it is difficult to fine-tune them directly, even resulting in catastrophic forgetting (Houlsby et al. 2019). Inspired by LoRA (Hu et al. 2021), adapter tuning with UT-Adapter (Guo et al. 2023) is employed. Assume that the feature representations of visual and audio modalities are $\mathbf{X}_{(v)}^t$ and $\mathbf{X}_{(a)}^t$ separately. When multiple modalities including both visual and audio are considered, we simply fuse them with average weights as $\mathbf{X}_{(f)}^t = 0.5\mathbf{X}_{(v)}^t + 0.5\mathbf{X}_{(a)}^t$. For simplicity, we use $\mathbf{X}^t$ as the target sample embeddings generated by $\mathcal{E}$ in the following, *i.e.*, $\mathbf{X}^t = \mathcal{E}(\mathbf{V}^t)$.

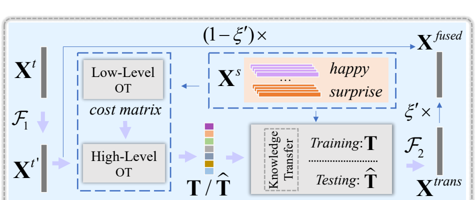


Figure 3: H-OTKT module. It formulates the relation mapping between source classes and target samples, and then performs knowledge transfer.

Hierarchical Optimal Transport Knowledge Transfer (H-OTKT) Module

As presented before, we aim to improve the classification performance by transferring VFER domain knowledge $\mathbf{X}^s$ to target deception domain. Since the label space and the distribution of two domains in the feature space are different, one of the key questions is how much knowledge of facial expression data should be transferred. Based on hierarchical optimal transport, we propose H-OT Knowledge Transfer (H-OTKT) module with high-level and low-level OT illustrated in Fig. 3.

In particular, high-level OT learn the optimal correlation between classes of VFER dataset and samples of deception dataset with a given cost matrix, where the cost matrix depends on the total low-level OT distance between each target deception sample and all samples from each class of VFER dataset.

$\mathbf{X}^t$ is firstly mapped into $\mathbf{X}^{t'} = \mathcal{F}_1(\mathbf{X}^t) \in \mathbb{R}^{n\times d}$ by an MLP $\mathcal{F}_1$, such that the feature spaces between source and target domain could be the same, where n is the batch size. Let $\mathcal{Q} = \sum_{k=1}^{L^s} \frac{1}{L^s}\delta_{\mathcal{Q}^k}$ as the discrete uniform distribution over L^s classes of VFER dataset, $\mathcal{Q}^k$ is the representation vector of k-th class. And $\mathcal{P} = \sum_{i=1}^{n} \frac{1}{n}\delta_{\mathbf{X}_i^{t'}}$ is the discrete uniform distribution over n target deception samples. Then, according to Eq. (2), the entropic regularized OT between $\mathcal{P}$ and $\mathcal{Q}$ is:

$$\mathrm{OT}_{high}(\mathcal{P},\mathcal{Q}) = \min_{\mathbf{T}\in\Pi(\mathcal{P},\mathcal{Q})} < \mathbf{T}, \mathbf{M} >_{\mathrm{F}} - \epsilon\mathcal{H}(\mathbf{T}) \tag{3}$$

where $\mathbf{T} \in \mathbb{R}^{n\times L^s}$ and $\mathbf{M} \in \mathbb{R}^{n\times L^s}$ are the transport plan and the cost matrix between facial expression classes and target deception samples. Each element $\mathbf{T}_{i,k}$ indicates the importance of the k-th class in VFER dataset for the i-th sample in deception mini-batch, determining which class and how much of knowledge should be transferred. Besides, $\mathbf{T}$ should satisfy the following constraint:

$$\Pi(\mathcal{P},\mathcal{Q}) := \left\{ \sum_{i=1}^{n} \mathbf{T}_{i,k} = \frac{1}{L^s},\ \sum_{k=1}^{L^s} \mathbf{T}_{i,k} = \frac{1}{n} \right\} \tag{4}$$

It is apparently that the solution $\mathbf{T}$ relies on the cost matrix $\mathbf{M}$, simply applying cosine similarity with the features of samples from deception mini-batch and the mean of features from each class of VFER dataset may lead to sub-optimal solution. Moreover, the contribution of different samples in each class may be various. So, we utilize another optimal transport formulation to obtain the optimal $\mathbf{M}$. According to (Guo et al. 2022), the empirical distribution of the k-th class is expressed as $\mathcal{Q}^k = \sum_{j=1}^{J_k} p_j^k \delta_{\mathbf{X}_j^{s,k}}$, where the importance p_j^k of the j-th sample in the k-th source class is obtained by logistic regression score. Therefore, a low-level entropic regularized OT is further defined as follows:

$$\begin{aligned}&\mathrm{OT}_{low}(\mathcal{P},\mathcal{Q}^k) = \\ &\min_{\mathbf{T}^{low,k}\in\Pi(\mathcal{P},\mathcal{Q}^k)} <\mathbf{T}^{low,k},\mathbf{M}^{low,k}>_{\mathrm{F}} -\epsilon\mathcal{H}(\mathbf{T}^{low,k})\end{aligned} \tag{5}$$

$\Pi(\mathcal{P},\mathcal{Q}^k) := \left\{ \sum_j^{J_k} \mathbf{T}_{i,j}^{low,k} p_j^k = \frac{1}{n}, \sum_i^n \mathbf{T}_{i,j}^{low,k} \frac{1}{n} = p_j^k \right\}$ is the constrain, and $\mathbf{T}^{low,k}$ is the transport plan between each sample in mini-batch and samples in the k-th source domain class. $\mathbf{M}^{low,k} \in \mathbb{R}^{n\times J_k}$ is determined by cosine similarity, *i.e.*, $\mathbf{M}_{i,j}^{low,k} = 1 - \cos(\mathbf{X}_i^{t'}, \mathbf{X}_j^{s,k})$. The cost matrix $\mathbf{M}$ in high-level OT of Eq. (3) will be replaced by the total OT distance between each target deception sample and all sample in each class of VFER dataset, *i.e.*, $\mathbf{M}_{:,k} =< \mathbf{T}^{low,k}, \mathbf{M}^{low,k} >_{\mathrm{F}}$.

For the optimization, both Eq. (5) and Eq. (3) are solved by Sinkhorn algorithm (Cuturi 2013) hierarchically. Using the OT distance calculated from low-level OT as the cost $\mathbf{M}$ of high-level OT adaptively, H-OTKT is able to obtain the transport weight $\mathbf{T}$ between deception samples and facial expression classes, which is the potential correlation mapping of facial expression classes for target samples.

Once we obtained correlation mapping $\mathbf{T}$ by solving Eq. (3), knowledge transformation can be performed. For each sample in deception domain, more knowledge from highly associated classes should be transferred, while knowledge from uncorrelated classes should not be transferred. To realize it, the transferred knowledge $\mathbf{X}^{trans} \in \mathbb{R}^{n\times d}$ is represented as follows:

$$\mathbf{X}_i^{trans} = \mathcal{F}_2\left(n \cdot \sum_{k=1}^{L^s} \mathbf{T}_{i,k} \left[\frac{1}{J_k} \sum_{j=1}^{J_k} \mathbf{X}_j^{s,k} \right] \right),\quad i = 1,\cdots,n \tag{6}$$

where $\frac{1}{J_k}\sum_{j=1}^{J_k} \mathbf{X}_j^{s,k}$ denotes the average feature of samples belonging to the k-th class in source domain; $\mathbf{T}_{i,k}$ quantifies the correlation weight between the k-th source class and i-th deception sample; n is used for scaling due to the constraint in high-level OT. And $\mathcal{F}_2$ is an MLP. In order to integrate the transferred knowledge $\mathbf{X}^{trans}$ with features $\mathbf{X}^{t'}$ extracted from target samples, the fused representation of deception detection samples are calculated as:

$$\mathbf{X}^{fused} = \xi'\mathbf{X}^{trans} + (1-\xi')\mathbf{X}^{t'} \tag{7}$$

where ξ' is the weight of transferred feature $\mathbf{X}^{trans}$. Since it's hard to learn excellent $\mathbf{X}^{t'}$ at the beginning of the training phase, a curriculum learning strategy (Kumar, Packer, and Koller 2010; Wang, Chen, and Zhu 2021) is adopted as $\xi' = \frac{\xi}{2} \times \left(1 - \cos\left(\frac{e-1}{N_e} \times \pi\right)\right)$, where e is the current training epoch number and N_e is the total training epoch number. As ξ' is gradually increased, a better $\mathbf{X}^{t'}$ is gained for H-OTKT.

Classification Layer

The final classification layer contains one MLP with softmax, which takes $\mathbf{X}^{fused}$ as input and outputs the predicted label $\hat{\boldsymbol{y}} \in \mathbb{R}^{n\times L^t}$:

$$\hat{\boldsymbol{y}} = \mathcal{F}_3(\mathbf{X}^{fused}) \tag{8}$$

Here, $\mathcal{F}_3$ is the MLP classifier. With ground truth label $\boldsymbol{y}^t = [y_1^t, \ldots, y_n^t]$, the classification loss function is formulated as:

$$\mathcal{L}_{ce}(\boldsymbol{y}^t, \hat{\boldsymbol{y}}) = -\mathbb{E}_{\boldsymbol{y}^t}[\log \hat{\boldsymbol{y}}] \tag{9}$$

where $\mathbb{E}$ is expectation. To reduce the difference between distribution spaces from source and target domain in H-OTKT, and further improve the final prediction, another loss function is defined based on the Sinkhorn divergence (Feydy et al. 2019) to obtain the space discrepancy between class average of $\mathbf{X}^s$ and $\mathbf{X}^{t'}$ (Nguyen and Luu 2022):

$$\mathcal{L}_{ot}(\mathbf{X}^{t'}, \mathbf{X}^s) = ds_{\mathrm{OT}}(\mathcal{P},\mathcal{Q}) - \frac{1}{2}ds_{\mathrm{OT}}(\mathcal{P},\mathcal{P}) - \frac{1}{2}ds_{\mathrm{OT}}(\mathcal{Q},\mathcal{Q}) \tag{10}$$

where $ds_{\mathrm{OT}}(\cdot,\cdot)$ is the total OT cost between two distributions solved by the regular OT (Eq. (1)) with cosine similarity as cost function. Then the total loss function is formulated as:

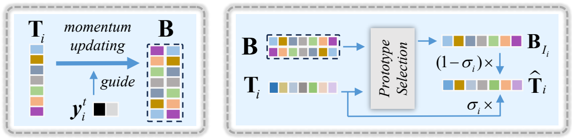
$$\mathcal{L} = \mathcal{L}_{ce} + \eta\mathcal{L}_{ot} \tag{11}$$

In Eq. (11), the $\mathcal{L}_{ce}$ term optimizes the whole network to improve the classification performance while the $\mathcal{L}_{ot}$ term is used for reducing the discrepance between the source feature space and the target feature space.

Sample-specific Re-weighting Knowledge Bank (SRKB) Module

Optimal knowledge from proper classes in facial expression dataset has been obtained by H-OTKT. Empirically, samples in deception dataset have varying semantics, thus the class relation would be various because of random data sampling (Wang et al. 2020). In order to more efficiently use learned knowledge from H-OTKT, and further improve the robustness during testing phase, a plug-in Sample-specific Re-weighting Knowledge Bank (SRKB) module with no additional trainable parameters shown in Fig. 4 (a) and (b) is constructed.

Aiming at obtaining more robust and general effective information from VFER dataset and eliminate the randomness caused by data sampling (Wang et al. 2020), a correlation prototype $\mathbf{B} = [\mathbf{B}_1^\top \ldots \mathbf{B}_{L^t}^\top]^\top \in \mathbb{R}^{L^t\times L^s}$ is constructed to store the robust category relation between two domains. $\mathbf{B}$ is initialized by $\frac{1}{L^s}$ to ensure $\sum_k^{L^s} \mathbf{B}_{l,k} = 1, l = 1,\ldots,L^t$. The l-th row of $\mathbf{B}$ demonstrates the association weights between the l-th class of target domain and each class of source



(a) Training Phase (b) Testing Phase

Figure 4: SRKB module. (a) Training phase: $\mathbf{B}$ is momentum updated to maintain the invariant knowledge of each target class relation with source classes; (b) Testing phase: SRKB module uses the learned $\mathbf{B}$ and sample-specific re-weighting strategy to enhance the detection performance.

domain. As shown in Fig. 4 (a), during training phase, momentum updating (Laine and Aila 2016) is introduced to update the correlation prototype $\mathbf{B}$ by:

$$\mathbf{B}_l = \alpha \mathbf{B}_l + (1-\alpha)\frac{1}{\sum_i^n \mathbb{I}_{\boldsymbol{y}_i^t=l}} \sum_{i=1}^{n} \mathbf{T}_i \mathbb{I}_{\boldsymbol{y}_i^t=l} \qquad (12)$$
$$l = 1, \cdots, L^t$$

where $\mathbb{I}_{\boldsymbol{y}_i^t=l} = 1$ if and only if the label of i-th deception sample $\boldsymbol{y}_i^t$ equals to l, otherwise $\mathbb{I}_{\boldsymbol{y}_i^t=l} = 0$. And α is the momentum factor. In Eq. (12), the relation between facial expression classes and each target categories (*i.e.*, truthful and deceptive) is accumulated into the correlation prototype $\mathbf{B}$ with the guidance of label $\boldsymbol{y}^t$. The underlying invariant relation between all deceptive samples and the source domain expressions is preserved.

According to the previous results of affective computing field (Rill-García et al. 2019) and psychology (DePaulo et al. 2003), deception should have high correspondences with some specific facial expression, such as *fear*, and *sad*. Ideally, for the i-th target deception sample, $\mathbf{T}_i$ should be sparse, where several elements are much higher than the others. This indicates the standard deviation of a sparse $\mathbf{T}_i$ should be larger if $\mathbf{T}_i$ is valid, otherwise it could be smaller. When an unsatisfactory transport plan $\mathbf{T}_i$ with small standard deviation is obtained, it is more reliable to use the corresponding correlation prototype $\mathbf{B}_{I_i}$ to quantify the correlation mapping instead, where I_i is the category that $\mathbf{T}_i$ may belong to. Otherwise, calculated $\mathbf{T}_i$ could be improved by its corresponding correlation prototype $\mathbf{B}_{I_i}$. In our experiment, $\mathbf{B}_{I_i}$ has smallest distance with $\mathbf{T}_i$ among $\{\mathbf{B}_1 \dots \mathbf{B}_{L^t}\}$. Mathematically, it can be formulated as:

$$\hat{\mathbf{T}}_i = \sigma_i \mathbf{T}_i + (1-\sigma_i)\mathbf{B}_{I_i}$$
$$s.t.\ \ \sigma_i = \begin{cases} 0, & \mathrm{std}(\mathbf{T}_i) < \nu \\ \mathrm{std}(\mathbf{T}_i) - \nu, & \text{otherwise} \end{cases} \qquad (13)$$

where $\mathrm{std}(\cdot)$ is standard deviation function, and ν is a threshold. When the standard deviation of $\mathbf{T}_i$ is greater than ν, we believe H-OTKT module finds a valid transport plan $\mathbf{T}$ to measure the importance between categories in source domain and samples in target domain. Otherwise, H-OTKT fails to find a valid transport plan due to noise, and $\mathbf{T}_i$ would not be used for transferring knowledge. Note that the standard deviation would always less than 1 according to Eq. (4). During the testing phase, $\mathbf{T}$ in Eq. (6) will be replaced by the obtained $\hat{\mathbf{T}} \in \mathbb{R}^{n \times L^s}$ in Eq. (13).

Experiments

Comparison Methods

We make comparisons with several deception detection methods on RTL (Pérez-Rosas et al. 2015) and DOLOS (Guo et al. 2023) in visual, audio and fused modalities to validate AFFAKT. Three *in-the-wild* VFER datasets are included, *i.e.*, DFEW (Jiang et al. 2020), FERV39K (Wang et al. 2022), MAFW (Liu et al. 2022). Refer to appendix for more details about experimental settings and datasets.

For machine learning based methods, visual (OpenFace and action units (AU)) and acoustic (MFCC and OpenSMILE) features are firstly selected, several widely-used classifier including SVM, Decision Tree (DT), Random Forest (RF), *etc.* are applied on visual features for classification and MLP is applied on acoustic features (Mathur and Matarić 2020; Avola et al. 2019; Yang, Liu, and C-H Huang 2021). For deep learning based approaches including ResNet18(RN18)+LSTM (Karnati et al. 2022; Ding et al. 2019; Guo et al. 2023), W2V2+MLP (Guo et al. 2023; Karnati et al. 2022; Krishnamurthy et al. 2018), RN18⊕OpenSMILE (Krishnamurthy et al. 2018; Guo et al. 2023) and PECL (Guo et al. 2023) are tested.

For transfer learning based methods, FreeLunch (Yang, Liu, and Xu 2021) and adaptive distribution calibration (ADC) (Guo et al. 2022) are adopted to obtain distance and the optimal transport plan between source class and target samples respectively, measuring the quantity $\mathbf{T}$ of knowledge from source classes to target samples. Then perform knowledge transfer via Eq. (6) and Eq. (7). PECL (Guo et al. 2023) is based on adapter-tuning, which is also a transfer learning based method. Another knowledge distillation based method Cr-KD-NCD (Gu et al. 2023) is also conducted, where the VFER classes are the known classes and deception classes are treated as novel classes. In this case, we employ the same visual encoder in AFFAKT as the backbone and only visual

Target		**RLT**			**DOLOS**		
Method	**Source**	**F1**	**ACC**	**AUC**	**F1**	**ACC**	**AUC**
OpenFace+SVM	-	0.2253	0.5293	0.5571	0.6975	0.5355	0.5430
OpenFace+DT	-	0.5553	0.5303	0.5303	0.5358	0.5058	0.5058
OpenFace+RF	-	0.6033	0.6033	0.5997	0.6175	0.5367	0.5466
OpenFace+AdaBoost	-	0.5199	0.5303	0.5766	0.5536	0.5057	0.5035
AU+SVM	-	0.4562	0.5043	0.4670	0.6813	0.5276	0.5242
AU+DT	-	0.4466	0.4643	0.4643	0.5453	0.5173	0.5173
AU+RF	-	0.5534	0.5463	0.5330	0.5808	0.5045	0.5157
AU+AdaBoost	-	0.5130	0.4877	0.4835	0.5295	0.4876	0.4735
OpenFace+LSTM	-	0.5241	0.5623	0.5952	0.5928	0.5628	0.5854
AU+LSTM	-	0.4888	0.6197	0.6760	0.6343	0.5646	0.5868
RN18+LSTM	-	0.4996	0.6117	0.6387	0.6415	0.5972	0.5668
PECL(only visual)	-	0.5880	0.6528	0.6734	0.7010	0.6387	0.6770
FreeLunch	DFEW	0.7612	0.8090	0.8712	0.6961	0.6222	0.6444
	FERV39K	0.7536	0.8173	0.8677	0.6831	0.6228	0.6456
	MAFW	0.7663	0.8173	0.8633	0.6695	0.6155	0.6459
ADC	DFEW	0.7793	0.8173	0.8674	0.6880	0.6716	0.7206
	FERV39K	0.7667	0.8173	0.8677	0.6830	0.6693	0.7156
	MAFW	0.7667	0.8173	0.8664	0.6938	0.6684	0.7180
Cr-KD-NCD	DFEW	0.6957	0.7200	0.6928	0.5850	0.6091	0.6013
	FERV39K	0.7805	0.7200	0.6464	0.6720	0.5879	0.5363
	MAFW	0.7778	0.6800	0.6368	0.7056	0.5697	0.5427
AFFAKT (ours)	DFEW	**0.8760**	**0.8670**	**0.8789**	0.7054	**0.6764**	**0.7212**
	FERV39K	0.8277	0.8340	0.8415	**0.7102**	0.6746	0.7203
	MAFW	0.8524	0.8500	0.8625	0.6948	0.6612	0.6970

Table 1: Results with visual modality.

Target		RLT			DOLOS		
Method	**Source**	**F1**	**ACC**	**AUC**	**F1**	**ACC**	**AUC**
MFCC+MLP	-	0.5226	0.6367	0.7030	0.5963	0.5810	0.6134
OpenSMILE+MLP	-	0.6885	0.6597	0.5926	0.6867	0.5537	0.5325
W2V2+MLP	-	0.6117	0.6780	0.6106	0.4383	0.5421	0.5369
PECL(only audio)	-	0.7121	0.7100	0.6962	0.6777	0.6119	0.6281
FreeLunch	DFEW	0.6396	0.6767	0.6869	0.6437	0.5864	0.6157
	FERV39K	0.6432	0.6850	0.6944	0.6589	0.5979	0.6194
	MAFW	0.6402	0.6767	0.6885	0.6490	0.5991	0.6196
ADC	DFEW	0.6402	0.6767	0.6858	0.6196	0.6058	0.6040
	FERV39K	0.6402	0.6767	0.6858	0.6165	0.6052	0.6039
	MAFW	0.6272	0.6767	0.6842	0.6129	0.6046	0.6039
AFFAKT (ours)	DFEW	0.7267	0.7270	0.7218	0.6822	0.6198	**0.6391**
	FERV39K	**0.7316**	0.7017	0.6917	**0.6982**	0.6173	0.6385
	MAFW	0.7266	**0.7440**	**0.7396**	0.6736	**0.6198**	0.6387

Table 2: Results with audio modality.

modality is evaluated. We refer readers to appendix for more details of these methods.

Comparison Results

Comparison results on both datasets with visual, audio and fused modality in terms of F1 score, ACC and AUC metrics are shown in Table 1. We only report the average value of different folds in the main text and the standard deviation between different folds are reported in the appendix.

Overall, our proposed method obtains the best performance across all evaluation metrics, which suggests that our proposed method is effective and advanced for deception detection. Besides, deep learning based models have achieved better results compared to machine learning based models. Such results indicate that deep learning model exhibits better feature extraction ability than traditional methods. Moreover, compared with (Guo et al. 2023), we can claim that extra facial expression knowledge is helpful to deception detection. AFFAKT has higher classification accuracy than transfer learning based methods (Yang, Liu, and Xu 2021; Guo et al. 2022), demonstrating that better knowledge of facial expression data could be transferred and leveraged in our method. On the other hand, compared to larger DOLOS dataset, AFFAKT outperforms deep learning based methods on smaller RLT more significantly, indicating that AFFAKT is able to show better detection performance on datasets with fewer samples, while other deep learning methods fail when given limited labeled deception data. The results vary across different facial expression datasets, since three facial expression datasets contain different expression categories, and different pre-trained encoders $\mathcal{G}$ are employed for each dataset, resulting in different representation abilities in VFER feature representation $\mathbf{X}^s$.

Target		RLT			DOLOS		
Method	**Source**	**F1**	**ACC**	**AUC**	**F1**	**ACC**	**AUC**
OpenFace⊕OpenSMILE	-	0.6895	0.6781	0.6212	0.6124	0.5986	0.5863
RN18⊕OpenSMILE	-	0.6283	0.6853	0.6598	0.5863	0.6152	0.6485
PECL	-	0.7102	0.6939	0.7424	0.7084	0.6597	0.6353
FreeLunch	DFEW	0.7473	0.8010	0.8497	0.6686	0.6258	0.6628
	FERV39K	0.7460	0.8010	0.8504	0.6504	0.6204	0.6574
	MAFW	0.7695	0.8093	0.8547	0.6807	0.6289	0.6669
ADC	DFEW	0.7493	0.8093	0.8446	0.6997	0.6746	0.7307
	FERV39K	0.7493	0.8093	0.8435	0.6819	0.6729	0.7295
	MAFW	0.7493	0.8010	0.8411	0.6976	0.6741	0.7274
AFFAKT (ours)	DFEW	0.8162	0.8180	0.8381	0.7073	**0.6810**	0.7226
	FERV39K	0.7946	0.8010	0.8357	**0.7149**	0.6774	**0.7289**
	MAFW	**0.8412**	**0.8427**	**0.8563**	0.7111	0.6780	0.7181

Table 3: Results with fused modalities.

In Table 2 and Table 3, we can find that AFFAKT achieves the best performance on both deception dataset in both audio and fused modalities. Specifically, for audio modality, the highest ACCs and F1 scores are achieved when the source domain is MAFW and FERV39K, respectively. The highest AUC is achieved on RLT and DOLOS when the source domains are MAFW and DFEW, respectively. For the fused modality, the best performance is achieved on RLT using MAFW as source domain. And on DOLOS, the best F1 score, ACC and AUC are obtained using FERV39K, DFEW and FERV39K as source domain. The discrepancy of source domain between different deception dataset and modalities would be caused by the robustness and the performance of pre-trained $\mathcal{G}$ as we have discussed above.

Ablation Studies

Ablation studies are conducted to verify H-OTKT and SRKB modules proposed in this paper. As our method contains four modules shown in Fig. 2: Encoder layer, H-OTKT module, classification layer, and SRKB module, the baseline method denoted as Case A only includes encoder layer and classification layer, where the classification layer is applied on extracted features $\mathbf{X}^t$ of deception data by encoder layer $\mathcal{E}$ directly. And the baseline model is only trained with classification loss Eq. (9).

Influence of H-OTKT module. To validate the effectiveness of our proposed H-OTKT module, we add the H-OTKT module based on the baseline method (Case B in Table 4). Case B contains encoder layer, H-OTKT module and classification layer, it is trained with total loss function Eq. (11). Recall that the H-OTKT captures the optimal relation mapping between VFER classes and deception samples, and performs knowledge transfer. Compared with the results of Case A, when DFEW is utilized as source dataset, the ACCs increase on almost all target datasets and modalities, which demonstrates that H-OTKT can facilitate deception detection accuracy by transferring knowledge from the source domain to the target deception domain. Moreover, when source dataset is FERV39K or MAFW, ACCs increase for RLT data with visual modality, while decrease a little for others, because of the different representative abilities in different pre-trained MAE-DFER encoder. Results in Case B indicate that H-

Case	Method ①	②	③	Source	Visual RLT	Visual DOLOS	Audio RLT	Audio DOLOS	Fused RLT	Fused DOLOS
A	✗	✗	✗	-	0.7840	0.6647	0.7100	0.6119	0.7843	0.6794
B	✓	✗	✗	DFEW	0.8093	0.6677	0.6937	0.6186	0.8013	0.6804
				FERV39K	0.8177	0.6646	0.6940	0.6143	0.7763	0.6610
				MAFW	0.8010	0.6307	0.6607	0.6046	0.7847	0.6540
C	✓	✓	✗	DFEW	0.8587	0.6738	0.7103	0.6173	0.8237	0.6727
				FERV39K	0.8340	0.6689	0.7080	0.6146	0.8021	0.6619
				MAFW	0.8417	0.6477	0.7357	0.6117	0.8427	0.6695
D	✓	✗	✓	DFEW	**0.8670**	**0.6764**	0.7270	0.6198	0.8180	**0.6810**
				FERV39K	0.8340	0.6746	0.7017	0.6173	0.8010	0.6774
				MAFW	0.8500	0.6612	**0.7440**	**0.6198**	**0.8427**	0.6780

Table 4: Ablation studies results. ① H-OTKT module, ② SRKB with *fixed* σ_i, ③ proposed SRKB.

OTKT lacks robustness for different deception datasets and different modalities.

Influence of SRKB module. To validate whether the proposed SRKB module is able to improve the robustness of AFFAKT, we add the SRKB module based on the Case B. This case corresponds to Case D in Table 4. As we have introduced in the previous section, SRKB is deemed to alleviate the randomness whilst training process, and store and more efficiently use the learned relation mapping by fine-tuning relation mapping during testing (Eq. (13)). From Table 4, ACCs obviously increase in Case D compared with the results in Case B on both deception datasets in all modalities. This improvement indicates that the SRKB module is able to boost the performance of AFFAKT. Specifically, when we bring our attention to the results between Case A, B and D on both datasets in visual modality, we are able to discover that SRKB effectively improves the robustness between different datasets. Comparing the results between Case A, B and D in audio and fused modalities, we can conclude that SRKB is able to improve the robustness in different modalities.

Influence of the sample-specific re-weighting strategy in SRKB. Furthermore, we also validate the effectiveness of our proposed sample-specific re-weighting strategy in SRKB, which is donated as Case C in Table 4. In this case, we *fix* the $\sigma_i = 0.2$ for each sample, which means SRKB would treat all deception samples equally when fine-tuning relation mapping. Compared with results in Case B, ACCs on RTL show improvements, but some decreases also appear on DOLOS in audio and fused modalities. Compared with Case D, the aforementioned decreases are eliminated. Such resutls illustrate that the sample-specific re-weighting strategy is able to automatically fine-tune the relation mapping to obtain better knowledge transfer guidance when noise in the test deception detection datasets leads to unsatisfactory relation mapping.

Influence of pre-trained encoder $\mathcal{G}$. To evaluate the effect of pre-trained encoder $\mathcal{G}$, we employ Former-DFER (Zhao and Liu 2021) as the pre-trained encoder $\mathcal{G}$ to extract the source features (the process in Fig. 2 (a)). Comparing the results of Table 5 with that of Table 1 to Table 3, ACCs on RTL and DOLOS datasets show better performance when the source features are extracted by MAE-DFER. This phenomenon demonstrates that better source feature space structure could facilitate knowledge transformation based on H-OT.

Interpretability Studies

In order to analyze the learned correlation prototype $\mathbf{B}$, we show the value of $\mathbf{B}$ in Table 6 when best accuracy is achieved on the two deception datasets by AFFAKT in visual modality. We refer readers to the appendix for the other modalities and their analysis. According to the results shown in Table 1, the

Modality	Visual		Audio		Fused	
Dataset	RLT	DOLOS	RLT	DOLOS	RLT	DOLOS
DFEW	0.8753	0.6751	0.7183	0.6199	0.8177	0.6728
FERV39K	0.8423	0.6711	0.7270	0.6102	0.7927	0.6649
MAFW	0.8587	0.6697	0.7020	0.6139	0.7927	0.6790

Table 5: Results when Former-DFER is employed as pre-trained encoder $\mathcal{G}$.

Dataset	Category	Source						
RLT	Target	happy	sad	neutral	angry	surprise	disgust	fear
	truthtul	0.3414	0.0999	0.1418	0.1394	0.1903	0.0546	0.0326
	deceptive	0.0318	0.4467	0.1097	0.1190	0.1600	0.0480	0.0847
	DIFF	**0.3096**	**0.3468**	0.0321	0.0204	0.0303	0.0066	**0.0521**
DOLOS		happy	sad	neutral	angry	surprise	disgust	fear
	truthtul	0.3908	0.0936	0.1688	0.0213	0.2343	0.0486	0.0426
	deceptive	0.1324	0.3508	0.0349	0.0923	0.2160	0.0943	0.0793
	DIFF	**0.2584**	**0.2572**	**0.1339**	0.0710	0.0183	0.0457	0.0367

Table 6: Learned $\mathbf{B}$ on RLT and DOLOS in visual modality. **DIFF** represents the absolute value of the difference between *truthful* and *deceptive*.

best ACCs are achieved for RLT and DOLOS when the source domain is selected as DFEW in viusal modality. Therefore, the corresponding results for both deception dataset in visual modality is shown in Table 6 when DFEW is employed as the source domain.

There are seven classes (*happy*, *sad*, *neutral*, *angry*, *surprise*, *disgust* and *fear*) in DFEW dataset (Jiang et al. 2020) and two classes (*truthful* and *deceptive*) in both deception datasets. Bold represents the largest three absolute difference (DIFF) between being truthful and deceptive across all categories of source domain in Table 6.

As demonstrated by Table 6, *sad* is significant related to *deceptive* for both RLT and DOLOS, while *happy* is remarkably correlated to *truthful*. Such results are also supported by psychological theory that liars will be less positive and pleasant than truth tellers (DePaulo et al. 2003; Zloteanu 2020; Hauch et al. 2015). For *neutral* and *fear*, *neutral* has a higher similarity with *truthful* than *deceptive* on both datasets, especially on DOLOS dataset. Besides, *fear* is more correlated with *deceptive* than *truthful* on both datasets. (Mathur and Matarić 2020) showed that deceivers in high-stakes situations are likely to associate with *fear*, which is consistent with our results.

Table 6 shows that AFFAKT can automatically establish proper relations between classes of facial expressions and deception, which helps in leveraging transferred knowledge from facial affective datasets in testing phase.

Conclusions

This paper presents a novel video deception detection method AFFAKT, aiming at addressing the challenge of insufficient high-quality large-scale labeled training datasets. We first develop H-OTKT module to perform knowledge transformation from related facial expression classes to deception samples, which estimates different weights of facial expression to deception samples by H-OT. Moreover, we design a correlation prototype based module SRKB to retain the invariant information within deceptive and truthful samples during training, which maintains the robust relation information between source and target classes. During testing phrase, SRKB integrates the predicted transport plan and the learned correlation prototype using a sample-specific re-weighting technique to leverage transferred knowledge. Extensive experiments have been conducted, showing that our proposed method outperforms other detection methods.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62306118 and No. 62207002), Basic and Applied Basic Research Foundation of Guangzhou (2023A04J1682), and the Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004), the Fundamental Research Funds for the Central Universities (2023ZYGXZR105).

References

Avola, D.; Cinque, L.; Foresti, G. L.; and Pannone, D. 2019. Automatic deception detection in rgb videos using facial action units. In *Proceedings of the 13th International Conference on Distributed Smart Cameras*, 1–6.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L.-P. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 59–66. IEEE.

Borza, D.; Itu, R.; and Danescu, R. 2018. In the Eye of the Deceiver: Analyzing Eye Movements as a Cue to Deception. *Journal of Imaging*, 4(10).

Chebbi, S.; and Jebara, S. B. 2023. Deception detection using multimodal fusion approaches. *Multimedia Tools and Applications*, 82(9): 13073–13102.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C. F.; and Huang, J.-B. 2019. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

DePaulo, B. M.; Lindsay, J. J.; Malone, B. E.; Muhlenbruck, L.; Charlton, K.; and Cooper, H. 2003. Cues to deception. *Psychological bulletin*, 129(1): 74.

Ding, M.; Zhao, A.; Lu, Z.; Xiang, T.; and Wen, J.-R. 2019. Face-Focused Cross-Stream Network for Deception Detection in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feydy, J.; Séjourné, T.; Vialard, F.-X.; Amari, S.-i.; Trouvé, A.; and Peyré, G. 2019. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2681–2690. PMLR.

Gu, P.; Zhang, C.; Xu, R.; and He, X. 2023. Class-relation knowledge distillation for novel class discovery. *lamp*, 12(15.0): 17–5.

Guo, D.; Tian, L.; Zhao, H.; Zhou, M.; and Zha, H. 2022. Adaptive distribution calibration for few-shot learning with hierarchical optimal transport. *Advances in Neural Information Processing Systems*, 35: 6996–7010.

Guo, X.; Selvaraj, N. M.; Yu, Z.; Kong, A. W.-K.; Shen, B.; and Kot, A. 2023. Audio-Visual Deception Detection: DOLOS Dataset and Parameter-Efficient Crossmodal Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22135–22145.

Hauch, V.; Blandón-Gitlin, I.; Masip, J.; and Sporer, S. L. 2015. Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personality and social psychology Review*, 19(4): 307–342.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2790–2799. PMLR.

Hsiao, S.-W.; and Sun, C.-Y. 2022. Attention-Aware Multimodal RNN for Deception Detection. In *2022 IEEE International Conference on Big Data (Big Data)*, 3593–3596. IEEE.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jaiswal, M.; Tabibu, S.; and Bajpai, R. 2016. The truth and nothing but the truth: Multimodal analysis for deception detection. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 938–943. IEEE.

Ji, Z.; Tian, X.; and Liu, Y. 2024. AFFAKT: A Hierarchical Optimal Transport Based Method for Affective Facial Knowledge Transfer in Video Deception Detection. arXiv:2412.08965.

Jiang, X.; Zong, Y.; Zheng, W.; Tang, C.; Xia, W.; Lu, C.; and Liu, J. 2020. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, 2881–2889.

Karnati, M.; Seal, A.; Yazidi, A.; and Krejcar, O. 2022. LieNet: A Deep Convolution Neural Network Framework for Detecting Deception. *IEEE Transactions on Cognitive and Developmental Systems*, 14(3): 971–984.

Krishnamurthy, G.; Majumder, N.; Poria, S.; and Cambria, E. 2018. A Deep Learning Approach for Multimodal Deception Detection. In *Conference on Intelligent Text Processing and Computational Linguistics*.

Kumar, M.; Packer, B.; and Koller, D. 2010. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23.

Laine, S.; and Aila, T. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Le, M. T. 2016. *Language of low-stakes and high-stakes deception: differences within individuals*. Ph.D. thesis, University of British Columbia.

Lee, J.; Dabagia, M.; Dyer, E.; and Rozell, C. 2019. Hierarchical optimal transport for multimodal distribution alignment. *Advances in neural information processing systems*, 32.

Liu, Y.; Dai, W.; Feng, C.; Wang, W.; Yin, G.; Zeng, J.; and Shan, S. 2022. MAFW: A Large-scale, Multi-modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, 24–32.

Mathur, L.; and Matarić, M. J. 2020. Introducing representations of facial affect in automated multimodal deception detection. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, 305–314.

Nguyen, T. T.; and Luu, A. T. 2022. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11103–11111.

Pérez-Rosas, V.; Abouelenien, M.; Mihalcea, R.; and Burzo, M. 2015. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 59–66.

Peyré, G.; and Cuturi, M. 2019. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6): 355–607.

Rill-García, R.; Jair Escalante, H.; Villasenor-Pineda, L.; and Reyes-Meza, V. 2019. High-level features for multimodal deception detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Şen, M. U.; Perez-Rosas, V.; Yanikoglu, B.; Abouelenien, M.; Burzo, M.; and Mihalcea, R. 2020. Multimodal deception detection using real-life trial data. *IEEE Transactions on Affective Computing*, 13(1): 306–319.

Snchez-Junquera, J.; Villaseor-Pineda, L.; y Gmez, M. M.; Rosso, P.; and Stamatatos, E. 2020. Masking domain-specific information for cross-domain deception detection. *Pattern Recognition Letters*, 135: 122–130.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H.; Xu, M.; Ni, B.; and Zhang, W. 2020. Learning to combine: Knowledge aggregation for multi-source domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, 727–744. Springer.

Wang, X.; Chen, Y.; and Zhu, W. 2021. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 4555–4576.

Wang, Y.; Sun, Y.; Huang, Y.; Liu, Z.; Gao, S.; Zhang, W.; Ge, W.; and Zhang, W. 2022. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20922–20931.

Yang, J.-T.; Liu, G.-M.; and C-H Huang, S. 2021. Multimodal Deception Detection in Videos via Analyzing Emotional State-based Feature. *arXiv e-prints*, arXiv–2104.

Yang, S.; Liu, L.; and Xu, M. 2021. Free Lunch for Few-shot Learning: Distribution Calibration. In *International Conference on Learning Representations*.

Zhang, H.; Ding, Y.; Cao, L.; Wang, X.; and Feng, L. 2022. Fine-Grained Question-Level Deception Detection via Graph-Based Learning and Cross-Modal Fusion. *IEEE Transactions on Information Forensics and Security*, 17: 2452–2467.

Zhao, Z.; and Liu, Q. 2021. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1553–1561.

Zloteanu, M. 2020. Reconsidering facial expressions and deception detection. *Handbook of facial expression of emotion*, 3: 238–284.

Zuckerman, M.; DePaulo, B. M.; and Rosenthal, R. 1981. Verbal and Nonverbal Communication of Deception. volume 14 of *Advances in Experimental Social Psychology*, 1–59. Academic Press.