Distribution Matching for Multi-Task Learning of Classification Tasks: A Large-Scale Study on Faces & Beyond

Dimitrios Kollias[1]*, Viktoriia Sharmanska[2,3], Stefanos Zafeiriou[3]
[1] School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[2]School of Engineering and Informatics, University of Sussex, UK
[3] Department of Computing, Imperial College London, UK
d.kollias@qmul.ac.uk, sharmanska.v@sussex.ac.uk, s.zafeiriou@imperial.ac.uk

Abstract

Multi-Task Learning (MTL) is a framework, where multiple related tasks are learned jointly and benefit from a shared representation space, or parameter transfer. To provide sufficient learning support, modern MTL uses annotated data with full, or sufficiently large overlap across tasks, i.e., each input sample is annotated for all, or most of the tasks. However, collecting such annotations is prohibitive in many real applications, and cannot benefit from datasets available for individual tasks. In this work, we challenge this setup and show that MTL can be successful with classification tasks with little, or non-overlapping annotations, or when there is big discrepancy in the size of labeled data per task. We explore task-relatedness for co-annotation and co-training, and propose a novel approach, where knowledge exchange is enabled between the tasks via distribution matching. To demonstrate the general applicability of our method, we conducted diverse case studies in the domains of affective computing, face recognition, species recognition, and shopping item classification using nine datasets. Our large-scale study of affective tasks for basic expression recognition and facial action unit detection illustrates that our approach is network agnostic and brings large performance improvements compared to the state-of-the-art in both tasks and across all studied databases. In all case studies, we show that co-training via task-relatedness is advantageous and prevents negative transfer (which occurs when MT model's performance is worse than that of at least one single-task model).

Introduction

Holistic frameworks, where several learning tasks are interconnected and explicable by the reference to the whole, are common in computer vision. A diverse set of examples includes a scene understanding framework that reasons about 3D object detection, semantic segmentation and depth reconstruction (Wang, Fidler, and Urtasun 2015), a face analysis framework that addresses face detection, landmark localization, gender recognition, age estimation (Ranjan et al. 2017), a universal network for low-, mid-, high-level vision (Kokkinos 2017), a large-scale framework of visual tasks for indoor scenes (Zamir et al. 2018). Most if not all prior works rely on building a multi-task framework where learning is done based on the ground truth annotations with full or partial overlap across tasks. During training, all the tasks are optimised simultaneously aiming at representation learning that supports a holistic view of the framework.

What differentiates our work from these holistic approaches is exploring the idea of task-relatedness as means for co-training different tasks. In our work, relatedness between tasks is either provided explicitly in a form of expert knowledge, or is inferred based on empirical studies. Importantly, in co-training, the related tasks exchange their predictions and iteratively teach each other so that predictors of all tasks can excel *even if we have limited or no data* for some of them. We propose an effective distribution matching and co-labeling approach based on distillation (Hinton, Vinyals, and Dean 2015), where knowledge exchange between tasks is enabled via distribution matching over their predictions.

Up until now training holistic models has been primarily addressed by combining multiple datasets to solve individual tasks (Ranjan et al. 2017), or by collecting the annotations in terms of all tasks (Zamir et al. 2018; Kokkinos 2017). For example, in affective computing, two most common tasks are predicting categorical expressions (e.g., happy, sad) and activation of action units (Ekman 1997) to explain the affective state. Collecting annotations of AUs is particularly costly, as it requires skilled annotators. The datasets collected so far (Li, Deng, and Du 2017; Benitez-Quiroz, Srinivasan, and Martinez 2016) have annotations for only one task and, despite significant effort, there is no dataset that has complete annotations of both tasks. Co-training via task relatedness is an effective way of aggregating knowledge across datasets and transferring it across tasks, especially with little or non-overlapping annotations, or when not many training data are available, or when there is a big discrepancy in the size of labeled data per task.

In this work we discuss two strategies to infer task-relatedness, via domain knowledge and dataset annotation. For example, the two aforementioned tasks of facial behavior analysis are interconnected with known strengths of relatedness in literature. In (Ekman 1997), the facial action coding system (FACS) was built to indicate for each of the basic expressions its *prototypical* AUs. In (Du, Tao, and Martinez 2014), a dedicated user study has been conducted to study the relationship between AUs and expressions. In (Khorrami, Paine, and Huang 2015), the authors show that

*Corresponding author

DNNs trained for expression recognition implicitly learn AUs. In our case study on face recognition, we have a dataset (CelebA (Liu et al. 2015)), where annotations for both tasks, identification and attribute prediction, are available. We can infer task relatedness empirically using its annotations.

One of the important challenges in MTL is how to avoid negative transfer, defined as when the performance of the multi-task model is worse than that of at least one single-task model (Wang et al. 2019; Liu, Liang, and Gitter 2019). Negative transfer occurs naturally in MTL scenarios when: i) source data are heterogeneous or less related (since tasks are diverse to each other, there is no suitable common latent representation and thus MTL produces poor representations); ii) one task or group of related tasks dominates the training process (negative transfer may occur simultaneously on tasks outside the dominant group).

To overcome negative transfer one can change the lambdas in the loss that control the importance of some tasks. However, it: could severely affect the performance on other tasks; is a computationally expensive procedure lasting days for each trial; is an ad-hoc method that is not guaranteed to work on other tasks or databases. To balance the performance across tasks, (Liu, Liang, and Gitter 2019) proposed a method that uses each task's training loss to indicate whether it is well trained, and then decreases the relative weights of the well trained tasks. The evaluation of performance indicators during each training iteration is costly. Negative transfer may be induced by conflicting gradients among the different tasks (Yu et al. 2020). (Lin et al. 2019) tackled this through multi-objective optimization, with decomposition of the problem into a set of constrained sub-problems with different trade-off preferences (among different tasks). However, this approach is rather complex, providing a finite set of solutions that do not always satisfy the MTL requirements and finally needs to perform trade-offs among tasks.

We demonstrate empirically that the proposed distribution matching and co-labeling approach based on task relatedness can prevent negative transfer in all our case studies. Via the proposed approach, knowledge of task relationship is infused in network training, providing it, in a simple manner, with higher level representation of the relationship between the tasks; it is not based on performance indicators and it does not perform any trade-offs between the different tasks. The main contributions of this paper are as follows:

- We propose a flexible approach that can accommodate different classification tasks by encoding prior knowledge of tasks relatedness. In our experiments we evaluate two effective strategies of task relatedness: a) obtained from domain knowledge and b) inferred empirically from dataset annotations (when domain knowledge is not available).
- We propose an effective weakly-supervised learning approach that couples, via distribution matching and label co-annotation, tasks with little, or even non-overlapping annotations, or with big discrepancy in their labeled data sizes; we consider a plethora of application scenarios, split in two case studies: i) affective computing; ii) beyond affective computing, including face recognition, fine-grained species categorization, shoe type classification, clothing categories recognition.
- We conduct an extensive experimental study utilizing 9 databases; we show that the proposed method is network agnostic (i.e., it can be incorporated and used in MTL networks) as it brings similar level of performance improvement in all utilized networks, for all tasks and databases. We also show that our method outperforms the state-of-the-art in all tasks and databases. Finally we show that our method successfully prevents negative transfer in MTL.

Related Work

Works exist in literature that use expression labels to complement missing AU labels or increase generalization of AU classifiers (Yang et al. 2016; Wang, Gan, and Ji 2017). Our work deviates from such methods (that only target AU detection), as we target joint learning of both behavior tasks via a single framework. In face analysis, the use of MTL is limited. In (Wang et al. 2017), MTL was tackled through a network that performed both facial recognition and attribute detection. A network was firstly trained for attribute detection and then used to generate predictions on a database with face identification labels. Then, a MT network was trained on that database. Network's loss was the sum of the independent task losses. In (Deng, Chen, and Shi 2020), a unified model for joint AU detection, expression recognition, and valence-arousal estimation was proposed; the utilized database contained images not annotated for all tasks. To tackle this, authors trained a teacher MT model with complete labels, then used it to generate predictions, which they used to train a student MT model; the student model outperformed the teacher one. The teacher model did not consider that tasks are interconnected - the overall loss was the sum of the independent task losses- and thus the student did not learn this relatedness. This work utilized only one database. In terms of MTL, (Sener and Koltun 2018) proposed MGDA-UB that casts MTL as multi-objective optimization to find a Pareto optimal solution, and proposed an upper bound for the loss. (Chen et al. 2018) proposed GradNorm, a gradient optimization algorithm that balances training in MTL by dynamically tuning gradient magnitudes. (Sun et al. 2020) proposed AdaShare, an adaptive sharing approach that decides what to share across which tasks by using a task-specific policy optimized jointly with the network weights. (Strezoski, Noord, and Worring 2019) proposed TR (Task Routing) that applies a conditional feature-wise transformation over the conv activations and is encapsulated in the conv layers.

The Proposed Approach

Let us consider a set of m classification tasks $\{\mathcal{T}_i\}_{i=1}^{m}$. In task $\mathcal{T}_i$, the observations are generated by the underlying distribution $\mathcal{D}_i$ over inputs $\mathcal{X}$ and their labels $\mathcal{Y}$ associated with the task. For the i-th task T_i, the training set D_i consists of n_i data points $(\mathbf{x}_j^i, y_j^i)$, $j = 1, \ldots, n_i$ with $\mathbf{x}_j^i \in \mathbb{R}^d$ and its corresponding output $y_j^i \in \{0, 1\}$ if it is a binary classification task, or $y_j^i \in \{0, 1\}^k$ (one-hot encoding) if it is a (mutually exclusive) k-class classification task.

The goal of MTL is to find m hypothesis: $h_1, ..., h_m$ over the hypothesis space $\mathcal{H}$ to control the average expected error over all tasks: $\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_i}\mathcal{L}(h_i(\mathbf{x}), y)$ with $\mathcal{L}$ being

	Cognitive Study (Du, Tao, and Martinez 2014)		Empirical Evidence, Aff-Wild2
Expression	Proto. AUs	Observational AUs (with weights w)	AUs (with weights w)
happiness	12, 25	6 (0.51)	12 (0.82), 25 (0.7), 6 (0.57), 7 (0.83), 10 (0.63)
sadness	4, 15	1 (0.6), 6 (0.5), 11 (0.26), 17 (0.67)	4 (0.53), 15 (0.42), 1 (0.31), 7 (0.13)
fear	1, 4, 20, 25	2 (0.57), 5 (0.63), 26 (0.33)	1 (0.52), 4 (0.4), 25 (0.85), 7 (0.57), 10 (0.57)
anger	4, 7, 24	10 (0.26), 17 (0.52), 23 (0.29)	4 (0.65), 7 (0.45), 25 (0.4), 10 (0.33)
surprise	1, 2, 25, 26	5 (0.66)	1 (0.38), 2 (0.37), 25 (0.85), 26 (0.3), 5 (0.5), 7 (0.2)
disgust	9, 10, 17	4 (0.31), 24 (0.26)	10 (0.85), 4 (0.6), 7 (0.75), 25 (0.8)

Table 1: Relatedness of expressions & AUs inferred from (Du, Tao, and Martinez 2014) (the weights denote fraction of annotators that observed the AU activation) or from Aff-Wild2 (the weights denote percentage of images that the AU was activated); 'Proto. AUs' stand for prototypical AUs

the loss function. We can also define a weight $\mathbf{w_i} \in \Delta^m$, $\{\mathbf{w}_i\}_{i=1}^m > 0$ to govern each task's contribution. The overall loss is: $\mathcal{L}_{MT} = \frac{1}{m}\sum_{i=1}^m \mathbf{w_i} \cdot \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_i}\mathcal{L}(h_i(\mathbf{x}), y)$.

In the following, we present the proposed framework via a plethora of case studies, mainly focusing on affective computing. The framework includes inferring the tasks' relationship and using it for coupling them during MTL. The coupling is achieved via the proposed co-annotation and distribution matching losses, which can be incorporated and used in any network that performs MTL, regardless of the input modality (visual, audio, text). The advantages of using these losses include: i) flexibility: no changes to network structure are made and no additional burden on inference is placed; ii) effectiveness: performance of various networks on multiple databases (small- or large-scale, image or video) is boosted and negative transfer is alleviated; iii) efficiency: negligible computational complexity is added during training; iv) easiness: a few lines of code are needed to implement.

Case Study I: Affective Computing

We start with the multi-task formulation of the behavior model. In this model we have two objectives: (1) learning 7 basic expressions, (2) detecting activations of 17 binary AUs. We train a multi-task model to jointly perform (1)-(2). For a given image $x \in \mathcal{X}$, we can have label annotations of either one of 7 basic expressions $y_{exp} \in \{1, 2, \ldots, 7\}$, or M AU activations $y_{au} \in \{0, 1\}^M$. For simplicity of presentation, we use the same notation x for all images leaving the context to be explained by the label notations. We train the multi-task model by minimizing the following objective:

$$\mathcal{L}_{MT} = \mathcal{L}_{Exp} + \mathcal{L}_{AU} + \mathcal{L}_{DM} + \mathcal{L}_{SCA} \tag{1}$$

where: $\mathcal{L}_{Exp} = \mathbb{E}_{x,y_{exp}}\big[-\log p(y_{exp}|x)\big]$ is cross entropy (CE) loss computed over images with basic expression label; $\mathcal{L}_{AU} = \mathbb{E}_{x,y_{au}}\big[-\log p(y_{au}|x)\big]$ is binary CE loss computed over images with M AU activations, with: log $p(y_{au}|x) := \frac{\sum_{i=1}^{M}\delta_i \cdot [y_{au}^i \log p(y_{au}^i|x) + (1-y_{au}^i)\log(1-p(y_{au}^i|x))]}{\sum_{k=1}^{M}\delta_k}$, $\delta_i \in \{0, 1\}$ indicates if the image contains AU_i annotation; $\mathcal{L}_{DM}$ and $\mathcal{L}_{SCA}$ are the proposed distribution matching and soft co-annotation losses, based on the relatedness of expressions and AUs; the losses' derivation is explained in the following.

Task-Relatedness

1) Obtained from Domain Knowledge: In the seminal work of (Du, Tao, and Martinez 2014), a cognitive and psychological study on the relationship between expressions and facial AU activations is conducted. The summary of the study is a Table of the relatedness between expressions and their prototypical and observational AUs, that we include in Table 1 for completeness. Prototypical AUs are ones that are labelled as activated across all annotators' responses; observational are AUs that are labelled as activated by a fraction of annotators.
2) Inferred Empirically from Dataset Annotations: If the above cognitive study is not available, we can infer task relatedness from external dataset annotations. In particular, we use the training set of Aff-Wild2 database (Kollias and Zafeiriou 2019, 2021a,b; Kollias et al. 2017, 2019, 2020b, 2023; Zafeiriou et al. 2017) to infer task relatedness, since this dataset is the first in-the-wild one that is fully annotated with basic expressions and AUs; this is shown in Table 1. In the following, we use as domain knowledge the cognitive and psychological study, to encode task relatedness and introduce the proposed approach for coupling the tasks.

Coupling of Basic Expressions and AUs

Via Distribution Matching Here, we propose the distribution matching loss for coupling the expression and AU tasks. The aim is to align the *predictions* of expression and AU tasks during training by making them consistent. From expression predictions we create new soft AU predictions and then match these with the network's actual AU predictions. For instance, if the network predicts *happy* with probability 1 and also predicts that AUs 4, 15 and 1 are activated, this is a mistake as these AUs are associated with the expression *sad* according to the prior knowledge. With this loss we infuse the prior knowledge into the network to guide the generation of better and consistent predictions.

For each sample x we have the predictions of expressions $p(y_{exp}|x)$ as the softmax scores over seven basic expressions and we have the prediction of AUs activations $p(y_{au}^i|x)$, $i = 1, \ldots, M$ as the sigmoid scores over M AUs. We match the distribution over AU predictions $p(y_{au}^i|x)$ with the distribution $q(y_{au}^i|x)$, where the AUs are modeled as a mixture over the basic expression categories:

$$q(y_{au}^i|x) = \sum_{y_{exp}\in\{1,\ldots,7\}} p(y_{exp}|x)\, p(y_{au}^i|y_{exp}), \tag{2}$$

where $p(y^i_{au}|y_{exp})$ is defined deterministically from Table 1 and is 1 for prototypical/observational action units, or 0 otherwise. For example, AU2 is prototypical for expression *surprise* and observational for expression *fear* and thus $q(y_{\text{AU2}}|x) = p(y_{\text{surprise}}|x)+p(y_{\text{fear}}|x)$. So with this matching if, e.g., the network predicts the expression *happy* with probability 1, i.e., $p(y_{\text{happy}}|x) = 1$, then the prototypical and observational AUs of *happy* -AUs 12, 25 and 6- need to be activated in the distribution q: $q(y_{\text{AU12}}|x) = 1$; $q(y_{\text{AU25}}|x) = 1$; $q(y_{\text{AU6}}|x) = 1$; $q(y^i_{au}|x) = 0,\ i \in \{1, .., 14\}$.

In spirit of the distillation approach, we match the distributions $p(y^i_{au}|x)$ and $q(y^i_{au}|x)$ by minimizing the cross entropy with the soft targets loss term, where all available train samples are used to match the predictions:

$$\mathcal{L}_{DM} = \mathbb{E}_x\left[\sum_{i=1}^{M}[-p(y^i_{au}|x)\log q(y^i_{au}|x)]\right] \tag{3}$$

Via Soft Co-annotation Here, we propose the soft co-annotation loss for coupling the expression and AU tasks. At first we create soft expression labels (that are guided by AU labels) by infusing prior knowledge of their relationship. Then we match these labels with the expression predictions. The new expression labels will help in cases of images with partial or no annotation overlap, especially if there are not many training data. We use the AU labels (instead of predictions) as they provide more confidence (the AU predictions -especially at the beginning of training- will be quite wrong; if we utilized this loss with the wrong AU predictions, it would also affect negatively the expression predictions).

Given an image x with ground truth AU annotations, y_{au}, we first co-annotate it with a *soft label* in form of the distribution over expressions and then match it with the predictions of expressions $p(y_{exp}|x)$. Thus, at first we compute, for each basic expression, an indicator score, $I(y_{exp}|x)$ over its prototypical and observational AUs being present:

$$I(y_{exp}|x) = \frac{\sum_{i\in\{1,\ldots,M\}} w^i_{au}\cdot y^i_{au}}{\sum_{i\in\{1,\ldots,M\}} w^i_{au}},\ y_{exp} \in \{1, \ldots, 7\} \tag{4}$$

where: w^i_{au} is 1 if AU i is prototypical for y_{exp} (from Table 1); is w if AU i is observational for y_{exp}; is 0 otherwise. For example, for expression *happy*, the indicator score $I(happy|x) = (y_{\text{AU12}} + y_{\text{AU25}} + 0.51\cdot y_{\text{AU6}})/(1+1+0.51)$.

Then, we convert the indicator scores to probability scores over expression categories; this *soft* expression label, $q(y_{exp}|x)$, is computed as following:

$$q(y_{exp}|x) = \frac{e^{I(y_{exp}|x)}}{\sum_{y'_{exp}} e^{I(y'_{exp}|x)}},\ \{y_{exp}, y'_{exp}\} \in \{1, .., 7\} \tag{5}$$

In this variant, every single image that has ground truth annotation of AUs will have a *soft* expression label assigned. Finally we match the predictions $p(y_{exp}|x)$ and the *soft* expression label $q(y_{exp}|x)$ by minimizing the cross entropy with the soft targets loss term:

$$\mathcal{L}_{SCA} = \mathbb{E}_x\left[\sum_{y_{exp}\in\{1,\ldots,7\}}[-p(y_{exp}|x)\log q(y_{exp}|x)]\right] \tag{6}$$

Case Study II: Beyond Affective Computing

Here, we show that our approach can also be used in other application scenarios: i) face recognition (facial attribute detection and face identification); ii) fine-grained species categorization (species classification and attribute detection); iii) shoe type recognition (shoe type classification and attribute detection); iv) clothing categories recognition (classification of clothing categories and attributes).

In the model's multi-task (MT) formulation, we have two objectives: (1) to detect M binary attributes, (2) to classify N classes. The aim of a MT model is to jointly perform (1) and (2). For a given image $x \in \mathcal{X}$, we can have labels of one of N classes $y_{cls} \in \{1, \ldots, N\}$, and M binary attributes $y_{att} \in \{0, 1\}^M$. We train the MT model by minimizing the objective: $\mathcal{L}_{MT} = \mathcal{L}_{Clc} + \mathcal{L}_{Att} + \mathcal{L}_{DM} + \mathcal{L}_{SCA}$, where:

$$L_{DM} = \mathbb{E}_x\left[\sum_{i=1}^{M}[-p(y^i_{att}|x)\log\sum_{y_{cls}} p(y_{cls}|x)p(y^i_{att}|y_{cls})]\right]$$

$$\mathcal{L}_{SCA} = \mathbb{E}_x\left[\sum_{y_{cls}}[-p(y_{cls}|x)\log\frac{e^{I(y_{cls}|x)}}{\sum_{y'_{cls}} e^{I(y'_{cls}|x)}}]\right]$$

$\mathcal{L}_{Cls}$ is the cross entropy loss for the classification task;
$\mathcal{L}_{att}$ is the binary cross entropy loss for the detection task;
$\mathcal{L}_{DM}$ is the distribution matching loss for matching the distributions $p(y^i_{att}|x)$ and the one where the attributes are modeled as a mixture over the classes;
$\mathcal{L}_{SCA}$ is the soft co-annotation loss for matching predictions $p(y_{cls}|x)$ and *soft* class labels (i.e., probability of each class indicator score, $I(y_{cls}|x)$, over its detected attributes);
$p(y^i_{att}|y_{cls}) = \frac{\text{total number of images with both } y^i_{att} \text{ and } y_{cls}}{\text{total number of images with } y_{cls}}$, is inferred empirically from dataset annotations.

Experimental Study

Databases: AffectNet (Mollahosseini, Hasani, and Mahoor 2017) with around 350K in-the-wild images annotated for 7 basic expressions (BE); **RAF-DB** (Li, Deng, and Du 2017) with around 15K in-the-wild images annotated for 7 BE; **ABAW4 LSD** (Kollias 2022a) -utilized in 4th Affective Behavior Analysis in-the-wild (ABAW) Competition at ECCV 2022- with around 280K in-the-wild synthetic images and 100K in-the-wild real images (which constitute the test set) annotated for 6 BE; **Aff-Wild2** (Kollias 2022b) -as utilized in 3rd ABAW Competition at CVPR 2022- with 564 in-the-wild videos (A/V) of around 2.8M frames annotated for 7 BE (plus 'other'), 12 AUs and valence-arousal; **EmotioNet** (Fabian Benitez-Quiroz, Srinivasan, and Martinez 2016) with around 50K images manually annotated for 11 AUs; **CelebA** with around 205K in-the-wild images of around 10.2K identities, each with 40 attributes (its split is subject independent; for our experiments, we generated

Databases	AffectNet - EmotioNet		RAF-DB - EmotioNet		ABAW4 LSD - EmotioNet		Aff-Wild2
Methods	EmoAffNet	EffNet-B2	PSR	VGGFACE	MTER-KDTD	HSE-NN	TMIF-FEA
Metrics	Acc - AFA	Acc - AFA	AA - AFA	AA - AFA	F1 - AFA	F1 - AFA	F1 - F1 (AU)
ST	66.4-71.9*	66.3-70.5*	80.8-69.6*	77.5-69.3*	35.9-70.0*	37.2-70.8*	35.9-49.9
NC MT	64.3*-75.1*	63.9*-73.8*	78.6*-72.8*	75.4*-72.2*	34.2*-72.2*	35.4-73.0*	33.7*-52.0*
S-T NC MT	64.9*-76.1*	64.5*-74.9*	79.8*-73.2*	76.5*-72.6*	35.0*-72.9*	36.1-73.7*	33.9*-52.2*
C MT (DM)	**69.4-80.0**	**68.9**-78.5	**84.8-78.1**	**81.4-77.2**	38.6 -77.5	39.3-78.3	38.8-55.9
C MT (DB)	**69.4-80.0**	68.7-**78.7**	83.8-77.2	80.5-76.0	**39.5-78.5**	**40.3-79.2**	**39.9-57.8**

Table 2: Performance comparison (in %) between various state-of-the-art single-task (ST) methods vs their multi-task counterparts with/without coupling (C MT/NC MT, respectively) under two relatedness scenarios (DB, i.e., Aff-Wild2, or DM, i.e., domain knowledge from (Du, Tao, and Martinez 2014)) vs their Student-Teacher (S-T) knowledge distillation counterparts; 'Acc': Accuracy; 'AA': Average Accuracy; 'AFA': average of F1 and mean Accuracy; * denotes our own implementation.

a new split into 3 subject dependent sets); **Caltech-UCSD Birds-200-2011** (Wah et al. 2011) (CUB) with around 12K images of 200 bird species and of 312 binary attributes; **Shoes** (Wah et al. 2011) (SADD) with around 15K women shoe images of 10 different types and of 10 attributes (Kovashka, Parikh, and Grauman 2012) (it does not have a predefined split, thus we split it in a training set of 7.4K and test set of 7.3K images). **Clothing Attributes Dataset** (Chen, Gallagher, and Girod 2012) (CAD) with around 2K images partially annotated for 7 clothing categories, 23 binary and 2 multi-class attributes (due to its very small size and to the non-predefined split, we perform 6 times 2-fold cross validation, i.e., we create 6 different 50-50 splits of the data).

Performance Measures: i) average accuracy for RAF-DB; ii) accuracy for AffectNet; iii) average of F1 and mean accuracy for EmotioNet; iv) F1 for ABAW4 LSD and Aff-Wild2; vii) accuracy and F1 for CelebA, CUB, SADD, CAD.

Pre-Processing & Training Implementation Details: Case Study I: with RetinaFace (Deng et al. 2020) extract bboxes & 5 landmarks (for alignment), resize images to $112 \times 112 \times 3$, Mixaugment (Psaroudakis and Kollias 2022) for augmentation. Case Study II: CelebA: use provided aligned images resized to $112 \times 112 \times 3$ (batch size = 200); CUB: use cropped images resized to $280 \times 280 \times 3$, label smoothing (value = 0.3), affine transformations (batch size = 150); SADD: resize images to $280 \times 280 \times 3$ (batch size = 100); CAD: resize images to $280 \times 280 \times 3$ (batch size = 50). In all databases, we used Adam, lr = 10^{-3} and images were normalized to $[-1, 1]$. Tesla V100 32GB GPU & Tensorflow were used.

Results on Case Study I: Affective Computing

Effectiveness of Proposed Coupling Losses Across Various Networks We utilized the state-of-the-art (sota) in various databases: i) AffectNet: EmoAfftNet(Ryumina, Dresvyanskiy, and Karpov 2022) and EffNet-B2 (Savchenko 2021); ii) RAF-DB: PSR (Vo et al. 2020) and VGGFACE (Kollias et al. 2020a); iii) ABAW4 LSD: HSE-NN (Savchenko 2022) and MTER-KDTD (Jeong et al. 2022); iv) Aff-Wild2: TMIF-FEA (Zhang et al. 2022). Let us note that EffNet-B2, MTER-KDTD and HSE-NN are multi-task methods.

The result of using each sota in single-task (ST) manner is shown in the row 'ST' of Table 2. The result of using each sota in MTL manner (e.g., EmoAfftNet trained on both AffectNet and EmotioNet; PSR trained on both RAF-DB and EmotioNet) is shown in row 'NC MT' (i.e., Multi-Task with no coupling) of Table 2. It might be argued that since more data are used for network training (i.e., the additional data coming from multiple tasks, even if they contain partial or non-overlapping annotations), the MTL performance will be better for all tasks. However, as shown and explained next, this is not the case as negative transfer can occur, or suboptimal models can be produced for some, or even all tasks (Wu, Zhang, and Ré 2019).

It can be seen in Table 2 (rows 'ST' and 'NC MT'), for all databases, that the sota, when trained in a MTL manner (without coupling), display a better performance for AU detection, but an inferior one for expression recognition - when compared to the corresponding performance of the single-task sota. *This indicates that negative transfer occurs in the case of basic expressions*. This negative transfer effect was due to the fact that the AU detection task dominated the training process. In fact, the EmotioNet database has a larger size than the RAF-DB, AffectNet and ABAW4 LSD. Negative transfer largely depends on the size of labeled data per task (Wang et al. 2019), which has a direct effect on the feasibility and reliability of discovering shared regularities between the joint distributions of the tasks in MTL.

Finally, we trained each of the sota networks in a MTL manner with the proposed coupling, under two relatedness scenarios; when the relatedness between the expressions and AUs was derived from the cognitive and psychological study of (Du, Tao, and Martinez 2014), or from dataset annotations (from Aff-Wild2 database). The former case is shown in row 'C MT (DM))' of Table 2 and the latter case in row 'C MT (DB)'. From Table 2 two observations can be made.

Firstly, when the proposed coupling is conducted, in each sota multi-task network, negative transfer is alleviated; the performance of all multi-task networks is better than the corresponding one of the single-task counterparts for both tasks. This is consistently observed in all utilized databases and experiments. Secondly, the use of the proposed coupling brings similar levels of performance improvement in all sota multi-task networks across the databases. In more detail, when coupling is conducted, networks outperform their counterparts without coupling by approximately: i) 5%

on Affectnet and 5% on EmotioNet (both EmoAfftNet and EffNet-B2); ii) 6% on RAF-DB and 5% on EmotioNet (both PSR and VGGFACE); iii) 5% on ABAW4 LSD and 6% on EmotioNet (both MTER-KDTD and HSE-NN); iv) 6% on Aff-Wild2 (MTER-KDTD).

To sum up, the use of coupling makes the MT networks greatly outperform their MT (without coupling) and single-task counterparts. This proves that the proposed coupling losses are network and modality agnostic as they can be applied and be effective in different networks and different modalities (visual, audio, A/V and text; e.g. TMIF-FEA is a multi-modal approach). This stands no matter which task relatedness scenario has been used for coupling the two tasks.

Finally, for comparison purposes we also used the Student-Teacher (S-T) knowledge distillation approach. We used one, or multiple teacher networks to create soft-labels for the databases that contain annotations only for one task, so that they contain complete, overlapping annotations for both tasks; we then trained a multi-task network on them. To illustrate this via an example: we use the single-task EmoAfftNet trained on AffectNet for expression recognition and test it on EmotioNet to create soft-expression labels; thus EmotioNet contains its AU labels and soft expression labels, i.e., the predictions of EmoAfftNet. Then we use the single-task EmoAfftNet trained on EmotioNet for AU detection and test it on AffectNet to create soft-AU labels; thus AffectNet contains its expression labels and soft AU labels. Then we train EmoAfftNet for MTL using both databases. We compare its performance to EmoAfftNet trained for MTL with the proposed coupling losses on both databases with their original non-overlapping annotations.

The results of the S-T approach are denoted in row 'S-T NC MT' (denoting Student-Teacher Multi-Task with no coupling) of Table 2. It can be observed that this approach shows a slightly better performance in both tasks compared to the multi-task counterparts that have been trained with the original non-overlapping annotations -without coupling-. This is expected as these networks have been trained with more annotations for both tasks. *Nevertheless, negative transfer for the basic expressions still occurs*. Moreover, our proposed approach greatly outperforms the S-T one. So overall, our proposed approach alleviates negative transfer and also brings bigger performance gain than that of S-T approach.

TMIF-FEA	Aff-Wild2	
	F1 (Expr)	F1 (AU)
no coupling	33.7	52.0
soft co-annotation (DM)	37.7	54.4
distr-matching (DM)	37.4	54.7
both (DM)	38.8	55.9
soft co-annotation (DB)	38.8	56.1
distr-matching (DB)	38.3	56.5
both (DB)	**39.9**	**57.8**

Table 3: Ablation Study on TMIF-FEA with/without coupling, under 2 relatedness scenarios; DM is domain knowledge from (Du, Tao, and Martinez 2014); DB is Aff-Wild2.

Ablation Study Here we perform an ablation study, utilizing the Aff-Wild2, on the effect of each proposed coupling loss on the performance of TMIF-FEA. Table 3 shows the results when task relatedness was drawn from domain knowledge, or from the training set of Aff-Wild2. It can be seen that when TMIF-FEA was trained with either or both coupling losses under any relatedness scenario, its performance was superior to the case when no coupling loss has been used. Finally, in both relatedness scenarios, best results have been achieved when TMIF-FEA was trained with both soft co-annotation and distr-matching losses. Similar results are yielded when we utilize each of the rest state-of-the-art on other databases, as explained in the previous subsection.

Results on Case II: Beyond Affective Computing

Effectiveness of Proposed Coupling Losses Across Broadly Used Networks Here, we utilized VGG-16, ResNet-50 and DenseNet-121. At first, we trained each of them for each application scenario for single-task learning (independent learning of classes and attributes) on CelebA, CUB, SADD and CAD datasets. In Table 4, these are denoted as '(2 ×) Single-Task'. We further trained these networks in MTL setting in two different cases: with coupling and without coupling during training. The presented results show the effectiveness of the proposed coupling losses to: i) avoid strong or mild negative transfer; ii) boost the performance of the multi-task models. The proposed coupling losses are network agnostic, as they bring similar level of improvement in all utilized networks, tasks and databases.

Face Recognition & Fine-Grained Species Categorization Table 4 shows that when the MTL baselines were trained without coupling, they displayed a better performance than the two single-task networks; this occurred in all studied cases, tasks, metrics and baseline models. This shows that the studied tasks were coherently correlated; training the multi-task architecture therefore, led to improved performance and no negative transfer occurred. Table 4 further shows that when the baselines were trained in the multi-task setting with coupling, they greatly outperformed its counterpart trained without coupling, in all studied tasks and metrics and for all baseline models. More precisely, when training with coupling, performance increased by at least: 4.3% and 5.9% in Accuracy and F1 Score for identity classification; 5% and 4.8% in Accuracy and F1 Score for species categorization; 1.2% and 2.3% in Accuracy and F1 Score for facial attribute detection; and 2% and 10.6% in Accuracy and F1 Score for species attribute detection.

Shoe Type Recognition Table 4 shows that negative transfer occurs in the case of attribute detection. Each single-task baseline for attribute detection displayed a better performance than its multi-task counterpart without coupling, whereas the latter displayed a better performance for shoe type classification. When the multi-task baseline networks were trained with coupling, the performance on both tasks was boosted and outperformed single- and multi-task counterparts, by at least 4.5% and 4.2% for classification and 2.1% and 2.7% for detection in both metrics. The proposed coupling losses overcame the negative transfer that had occurred.

	Db	Setting											
		2× Single-Task				Multi-Task with no Coupling				Multi-Task with Coupling			
		Classes		Attributes		Classes		Attributes		Classes		Attributes	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
V	CelebA	78.1	70.0	87.6	67.9	80.8	72.0	89.4	68.7	**85.9**	**78.0**	**90.6**	**71.0**
	CUB	78.2	78.4	85.2	27.0	80.0	80.2	85.5	28.6	**85.1**	**85.1**	**88.0**	**39.3**
	SADD	71.9	71.2	91.1	89.2	72.2	72.0	90.4	88.5	**76.4**	**76.5**	**93.4**	**91.3**
	CAD	52±5	38±7	80±2	40±2	41±7	32±9	75±3	33±4	**64±3**	**52±6**	**85±1**	**46±1**
R	CelebA	80.8	72.9	90.1	71.4	84.0	75.1	92.0	72.3	**88.6**	**81.1**	**93.3**	**74.7**
	CUB	82.8	82.8	89.5	30.8	84.3	84.3	89.9	32.5	**89.3**	**89.5**	**92.1**	**43.3**
	SADD	74.7	74.6	92.6	90.6	75.0	75.1	92.0	90.1	**79.3**	**79.4**	**94.7**	**93.2**
	CAD	55±5	41±7	84±2	44±2	44±7	35±9	79±3	38±4	**67±3**	**55±6**	**89±1**	**50± 1**
D	CelebA	80.1	72.2	89.9	70.2	82.8	74.1	91.7	71.0	**87.1**	**80.0**	**93.0**	**73.3**
	CUB	80.6	80.7	87.6	29.1	82.0	82.5	88.0	30.9	**87.2**	**87.3**	**90.0**	**41.5**
	SADD	73.3	73.1	91.3	89.6	73.7	73.7	90.7	88.9	**78.0**	**78.0**	**93.8**	**92.4**
	CAD	53±5	39±7	82±2	42±2	42±7	33±9	77±3	36±4	**65±3**	**53±6**	**87±1**	**48±1**

Table 4: Performance evaluation (in %) on various databases by three widely used baseline networks, VGG-16 (denoted as 'V'), ResNet-50 (denoted as 'R') and DenseNet-121 (denoted as 'D'); 'Acc' denotes accuracy; 'Db' denotes Database

	NTS-ST	NTS-NC MT	NTS-C MT	MGDA-UB	GradNorm	AdaShare	TR
CUB	87.5-92.0	89.6-92.7	94.4-95.2	86.3-90.9	86.0-90.4	86.2-90.6	83.2-76.5
SADD	77.7-93.0	78.0-92.1	82.6-94.9	76.5-93.1	76.1-92.8	76.7-93.4	73.1-78.7
CAD	60±5-86±2	49±6 - 81±4	64±2 - 92±1	52±4 - 83±2	51±4 - 82±3	49±4 - 79±3	61±5 - 74±3

Table 5: Accuracy evaluation (in %; in form Classes-Attributes) on various databases vs state-of-the-art and MTL methods

Clothing Categories Recognition Table 4 presents the outcomes of the 2-fold cross validation experiments (performed 6 times) in which the results are averaged and their spread is also shown (in the form: mean ± spread). From Table 4, it can be seen that the selected tasks are very heterogeneous and less correlated as all multi-task baselines without coupling performed significantly worse than single-task counterparts in all utilized metrics. Such severe negative transfers occurs as there is a big discrepancy in the size of labeled data per task in CAD dataset (the missing values for each attribute range from 12% to 84%) and its size is very small (it contains only 1856 images). When the multi-task baselines were trained with coupling, negative transfer was prevented and the models significantly outperformed their single-task counterparts (10-14% difference in Total Accuracy and 13-15% in F1 Score for classification; 4-5% in Total Accuracy and 5-7% in F1 Score for attributes). Finally, a smaller spread of the results can be observed in the case when the models were trained with coupling.

Effectiveness of Proposed Coupling Losses Across the State-of-the-Art At first, we show that the proposed coupling losses can also be incorporated in sota networks and thus we implement *NTS-Net* (Yang et al. 2018) in single task setting (denoted NTS-ST), in MTL setting without coupling (NTS-MT NC) and in MTL setting by adding our proposed coupling losses (NTS-MT C). Results are shown on Table 5 and are in accordance with the previous presented results (similar performance gain and alleviation of negative transfer). We then compare our method against MTL ones -presented in related work section- and thus we implement (ResNet50): MGDA-UB, GradNorm, AdaShare and TR. Table 5 presents their results. Comparing these with ResNet50 C MT of Table 4, it is evident that our method significantly outperforms all of them. Also, when comparing them to ST ResNet of Table 4: i) MGDA-UB, GradNorm and AdaShare cannot alleviate negative transfer in CAD for both tasks; ii) TR cannot alleviate negative transfer in CUB and CAD for attribute detection and in SADD for both tasks.

Conclusion & Limitation

We proposed a method for accommodating classification tasks by encoding prior knowledge of their relatedness. Our method is important as deep neural networks cannot necessarily capture tasks' relationship, especially in cases where: i) there is no or partial annotation overlap between tasks; ii) not many training data exist; iii) one task dominates the training process; iv) sub-optimal models for some tasks are produced; vi) there is big discrepancy in the size of labeled data per task. We considered a plethora of application scenarios and conducted extensive experimental studies. In all experiments our method helped the MT models greatly improve their performance compared to ST and MT models without coupling. Our method further helped alleviate mild or significant negative transfer that occurred when MT models displayed worse performance in some or all studied tasks than ST models. Our approach is general and flexible as long as there is a direct relationship between the studied tasks; the latter is our method's requirement and thus its limitation.

References

Benitez-Quiroz, C.; Srinivasan, R.; and Martinez, A. 2016. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR'16)*. Las Vegas, NV, USA.

Chen, H.; Gallagher, A.; and Girod, B. 2012. Describing clothing by semantic attributes. In *European conference on computer vision*, 609–623. Springer.

Chen, Z.; Badrinarayanan, V.; Lee, C.-Y.; and Rabinovich, A. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, 794–803. PMLR.

Deng, D.; Chen, Z.; and Shi, B. E. 2020. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, 828–835. IEEE Computer Society.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5203–5212.

Du, S.; Tao, Y.; and Martinez, A. M. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15): E1454–E1462.

Ekman, R. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

Fabian Benitez-Quiroz, C.; Srinivasan, R.; and Martinez, A. M. 2016. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5562–5570.

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

Jeong, J.-Y.; Hong, Y.-G.; Oh, J.; Hong, S.; Jeong, J.-W.; and Jung, Y. 2022. Learning from Synthetic Data: Facial Expression Classification based on Ensemble of Multi-task Networks. *arXiv preprint arXiv:2207.10025*.

Khorrami, P.; Paine, T.; and Huang, T. 2015. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 19–27.

Kokkinos, I. 2017. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6129–6138.

Kollias, D. 2022a. ABAW: learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, 157–172. Springer.

Kollias, D. 2022b. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2328–2336.

Kollias, D.; Cheng, S.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020a. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, 128(5): 1455–1484.

Kollias, D.; Nicolaou, M. A.; Kotsia, I.; Zhao, G.; and Zafeiriou, S. 2017. Recognition of affect in the wild using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 26–33.

Kollias, D.; Schulc, A.; Hajiyev, E.; and Zafeiriou, S. 2020b. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 637–643. IEEE.

Kollias, D.; Tzirakis, P.; Baird, A.; Cowen, A.; and Zafeiriou, S. 2023. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5888–5897.

Kollias, D.; Tzirakis, P.; Nicolaou, M. A.; Papaioannou, A.; Zhao, G.; Schuller, B.; Kotsia, I.; and Zafeiriou, S. 2019. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 1–23.

Kollias, D.; and Zafeiriou, S. 2019. Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace. *arXiv preprint arXiv:1910.04855*.

Kollias, D.; and Zafeiriou, S. 2021a. Affect Analysis in-the-wild: Valence-Arousal, Expressions, Action Units and a Unified Framework. *arXiv preprint arXiv:2103.15792*.

Kollias, D.; and Zafeiriou, S. 2021b. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3652–3660.

Kovashka, A.; Parikh, D.; and Grauman, K. 2012. Whittlesearch: Image search with relative attribute feedback. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2973–2980. IEEE.

Li, S.; Deng, W.; and Du, J. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2852–2861.

Lin, X.; Zhen, H.-L.; Li, Z.; Zhang, Q.; and Kwong, S. 2019. Pareto Multi-Task Learning. In *Thirty-third Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Liu, S.; Liang, Y.; and Gitter, A. 2019. Loss-balanced task weighting to reduce negative transfer in multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9977–9978.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*.

Psaroudakis, A.; and Kollias, D. 2022. Mixaugment & mixup: Augmentation methods for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2367–2375.

Ranjan, R.; Sankaranarayanan, S.; Castillo, C. D.; and Chellappa, R. 2017. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 17–24. IEEE.

Ryumina, E.; Dresvyanskiy, D.; and Karpov, A. 2022. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing*, 514: 435–450.

Savchenko, A. V. 2021. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*, 119–124. IEEE.

Savchenko, A. V. 2022. HSE-NN Team at the 4th ABAW Competition: Multi-task Emotion Recognition and Learning from Synthetic Images. *arXiv preprint arXiv:2207.09508*.

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. *arXiv preprint arXiv:1810.04650*.

Strezoski, G.; Noord, N. v.; and Worring, M. 2019. Many task learning with task routing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1375–1384.

Sun, X.; Panda, R.; Feris, R.; and Saenko, K. 2020. Adashare: Learning what to share for efficient deep multi-task learning. *Advances in Neural Information Processing Systems*, 33: 8728–8740.

Vo, T.-H.; Lee, G.-S.; Yang, H.-J.; and Kim, S.-H. 2020. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access*, 8: 131988–132001.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wang, S.; Fidler, S.; and Urtasun, R. 2015. Holistic 3d scene understanding from a single geo-tagged image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3964–3972.

Wang, S.; Gan, Q.; and Ji, Q. 2017. Expression-assisted facial action unit recognition under incomplete AU annotation. *Pattern Recognition*, 61: 78–91.

Wang, Z.; Dai, Z.; Póczos, B.; and Carbonell, J. 2019. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11293–11302.

Wang, Z.; He, K.; Fu, Y.; Feng, R.; Jiang, Y.-G.; and Xue, X. 2017. Multi-task deep neural network for joint face recognition and facial attribute prediction. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, 365–374. ACM.

Wu, S.; Zhang, H. R.; and Ré, C. 2019. Understanding and Improving Information Transfer in Multi-Task Learning. In *International Conference on Learning Representations*.

Yang, J.; Wu, S.; Wang, S.; and Ji, Q. 2016. Multiple facial action unit recognition enhanced by facial expressions. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 4089–4094. IEEE.

Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; and Wang, L. 2018. Learning to navigate for fine-grained classification. In *Proceedings of the European conference on computer vision (ECCV)*, 420–435.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient Surgery for Multi-Task Learning. *Advances in Neural Information Processing Systems*, 33.

Zafeiriou, S.; Kollias, D.; Nicolaou, M. A.; Papaioannou, A.; Zhao, G.; and Kotsia, I. 2017. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, 1980–1987. IEEE.

Zamir, A. R.; Sax, A.; Shen, W.; Guibas, L. J.; Malik, J.; and Savarese, S. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3712–3722.

Zhang, W.; Qiu, F.; Wang, S.; Zeng, H.; Zhang, Z.; An, R.; Ma, B.; and Ding, Y. 2022. Transformer-based Multimodal Information Fusion for Facial Expression Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2428–2437.