Towards a General Model of Knowledge for Facial Analysis by Multi-Source Transfer Learning

Valentin Vielzeuf[1,2] Alexis Lechervy[2] Stéphane Pateux[1] Frederic Jurie[2]

[1]Orange Labs, Cesson-Sévigné, France
[2]Université Caen Normandie, France

Abstract

This paper proposes a step toward obtaining general models of knowledge for facial analysis, by addressing the question of multi-source transfer learning. More precisely, the proposed approach consists in two successive training steps: the first one consists in applying a combination operator to define a common embedding for the multiple sources materialized by different existing trained models. The proposed operator relies on an auto-encoder, trained on a large dataset, efficient both in terms of compression ratio and transfer learning performance. In a second step we exploit a distillation approach to obtain a lightweight student model mimicking the collection of the fused existing models. This model outperforms its teacher on novel tasks, achieving results on par with state-of-the-art methods on 15 facial analysis tasks (and domains), at an affordable training cost. Moreover, this student has 75 times less parameters than the original teacher and can be applied to a variety of novel face-related tasks.

1. Introduction

An increasing number of deep neural networks has been implemented and trained during the last few years. These existing models can be seen as incredibly rich and compressed sources of knowledge about diverse domains, which can be reused to tackle novel tasks by transferring this knowledge. In this context, the standard way for knowledge transfer consists in selecting a single source, generally under the form of the parameters of a statistical model (*e.g.*, a pre-trained network), and to re-use it on a new task by fine-tuning the parameters. The knowledge source is often empirically chosen, typically by selecting the closest task according to human judgment or a complex rich task such as the ImageNet classification task.

To automate this selection process, recent works [40, 2, 67] have shown that a relational space between diverse basic tasks/models can be exploited, allowing to combine few potential candidate models and make the transfer more informative than using a single model.

However, the models discarded by this selection process may still contain useful knowledge. By analogy with multimodal problems, one may consider each independent model (specifically the last hidden representation of the neural networks) as a modality for the new task. Some modalities taken in isolation can yield bad results while they carry useful information complementary to other modalities. On the contrary, two well-performing modalities can be redundant and combining them brings less improvement.

Extending this reasoning to M modalities can be done by learning a common representation embracing all the modalities. However, this is not a trivial problem. For example, naively concatenating the local embeddings provided by each model does indeed produce a common embedding, but does not work well when dealing with many tasks/modalities, as observed by Zamir *et al.* [67].

We can formulate this multi-source transfer learning as follows. We define knowledge items as (domain, task) pairs, which can be associated with a model trained on this domain to fulfill this task. Supposing having M of such source knowledge, how to group them into a unique and general model of knowledge, performing well both on source knowledge and on novel target knowledge?

By addressing such a goal, this paper makes three main contributions: **(a)** the definition of a carefully designed ensemble of source and target knowledge related to facial analysis. Facial analysis is a topic of broad interest having received a lot of attention from the community, and for which large sets of knowledge (*i.e.* pre-trained models) are available. Moreover, a general knowledge in such a field would be of practical interest, as the number of new face-related tasks and domains grows exponentially. **(b)** a simple yet efficient methodology to project the source embeddings into a unique one, which is accurate both on the source and the target knowledge. **(c)** A distillation process is introduced to transfer the learned general knowledge into one lightweight convolutional neural network. This simple model outperformed its teacher, being on par with state-of-the-art models specifically built for solving specific tasks on specific domains. Moreover, it fits real-world application requirements, with 2 million parameters.

2. Related Work

The proposed work is related to several fields including information compression, transfer learning or distillation.

Information Fusion and Multi-Task Learning Combining the knowledge from several existing models into a single representation can be seen as a multimodal fusion, where model features are the input modalities. Classical methods exist such as Principal Components Analysis [35] or Canonical Correlation Analysis [30]. The recent trend is often focusing on equivalent neuronal methods such as the one Ngiam *et al*. [49] proposed, using multimodal autoencoders to learn a representation able to reconstruct all the modalities. The literature on deep multimodal fusion can be divided in i) multimodal architectures [62, 48, 53] focusing on where to fuse the information in the network, ii) representations based on constraints [3, 14, 61] building on the relations between the modalities (*e.g*. correlations). More details about this field can be found in these two surveys [5, 8].

Information fusion and multitask learning approaches are often related, as in the multitask autoencoder proposed by Ghifary *et al*. [26] allowing to improve domain generalization. The recent work by Ruder *et al*. [59] on multitasking architecture search allows to learn latent architectures for multitasking problems. Multitask approaches for face analysis have been proven to be as efficient or even better than single-task learning [15]. In the context of this paper, we can also see the fusion of diverse existing models both as a multimodal problem (each model is a modality) and as a multitask problem (the final student model has to be efficient on all tasks).

Transfer Learning As the goal of the paper is to transfer knowledge to novel tasks, it can be related to transfer learning. We share the same motivations as Taskonomy proposed by Zamir *et al*. [67], helping to select which combination of existing models to use when tackling a new task. Nevertheless in our case, because the existing models are already partially related to the target task, there is a benefit to select them all and keep what is useful in each model. Geyer *et al*. [25] propose to merge two pre-trained models before transferring them, based on incremental moment matching [39]. Chen *et al*. [17] designed a coupled end-to-end transfer learning, distilling the knowledge of one source model into the target model, while selective adversarial networks are proposed in [12] to select positive transfers and discard negative ones in the particular case where the target label space is a subspace of the source label space. Finally, regarding domain generalization, Mancini *et al*. [46] propose to fuse the outputs of domain-specific neural networks, after predicting the domain of target samples.

Self-supervised Learning Our work uses a large dataset to learn an autoencoder in an unsupervised fashion, and, by this means, exhibits the relations between the given models. Therefore, training the Teacher can be seen as a self-supervised learning process, which is a widely explored topic [69, 70, 51]. The goal of self-supervised learning is to design an efficient and cost-less proxy task helping to solve the target task for which we don't have enough annotations. Doersch *et al*. [23] showed the benefit of using several proxy tasks in self-supervised learning, which is basically what we are doing with the six different embeddings to reconstruct. The difference lies in the very definition of the proxy task. Ours are coming from previously learned knowledge. In traditional self-supervised, the tasks are low-level objectives such as solving a jigsaw puzzle [50] or evaluating the rotation of an image [27]. Radenovic *et al*. [54] used a proxy task closer to ours, using state-of-the-art models to extract edges of images as labels for their visual model.

Other methods falling in this unsupervised learning category can also be linked to ours, such as the deep clustering approach [13] where a convolutional neural network if trained using the output of a k-means clustering algorithm.

Model Compression and Knowledge Distillation A last important aspect of our method is to distill [32] the learned knowledge into a single lightweight model. In this spirit, Romero *et al*. proposed a deep and thin student named Fitnet [56]. It learns from both last layer and hidden layers of the teacher and outperforms it. Aiming to ensure privacy of the learning dataset, Papernot *et al*. [52] proposed to use several teachers in a semi-supervised fashion.

With a goal close to ours, Chebotar *et al*. [16] used as a teacher a weighted average prediction of an ensemble of neural networks. More recently the idea of data distillation was developed by [55], applying the same teacher model to diverse transformations of the input image and taking the average prediction as a label for the student. Li *et al*. [43] use feature map attention to regularize the learning of the student. Finally, multimodal distillation bears similarities with our approach, using some well-known modalities of a given input to master other views with fewer annotations. For instance, the SoundNet model [6] learns an audio model from the labels yield by a visual model, while Xu *et al*. [65] proposed to fuse the predictions coming from diverse modalities to improve the quality of the final main task of the student.

3. Methodology

To the best of our knowledge, transferring multi-source knowledge across both tasks and domains has not been addressed yet in the literature. Therefore, we will first study how to formally and practically formalize such a problem. We will then detail the two main steps composing our approach: dimensionality reduction and distillation.

3.1. Multi-source Multi-domain Transfer Learning

General formulation Let's first define the concept of *knowledge* as the abstract ability to perfectly solve a given task t on a given domain d. We limit ourselves to the family of tasks including classification / regression with machine learning techniques. The knowledge extracted when solving this problem (*e.g.* through the training of a deep neural network) can be denoted as $\mathcal{K}_{(t,d)} = (\mathcal{E}_{(t,d)}, \mathcal{C}_{(t,d)})$, where

- $\mathcal{E}_{(t,d)}$ is a function able to map each element x of domain d into a common embedding h
- $\mathcal{C}_{(t,d)}$ is a function able to map each embedding h to the expected output y

When dealing with a new target problem defined by a task t' on a domain d', $\mathcal{K}_{(t,d')}$ or $\mathcal{K}_{(t',d)}$ can be used to initialize a proposal for $\mathcal{K}_{(t',d')}$, that can be further adapted to fit the task/domain. *Transfer learning* consists in reusing (or fine-tuning) $\mathcal{E}_{(t,d)}$ and learning a new specific $\mathcal{C}_{(t',d)}$. While *domain adaptation* leads to learn $\mathcal{E}_{(t,d+d')}$ and $\mathcal{C}_{(t,d+d')}$.

The general problem we tackle is the one of adapting transfer learning and domain adaptation to the case where we have not only one already solved problem, but a set $\mathcal{S}$ of solved problems.

Following transfer learning and domain transfer, our approach is based on defining an operator $\mathcal{E}^G = \mathcal{G}(\mathcal{E}_{(t_i,d_i)}, i \in \mathcal{S})$ where $\mathcal{G}$ is a combination function using all the h_i embedding (coming from the $\mathcal{K}^S_{(t^S_i,d^S_i)}$). It aims at gathering all knowledge information regarding the t_i and the d_i into one embedding h_G.

One possible solution is to concatenate all the outputs of $\mathcal{E}_{(t_i,d_i)}, i \in \mathcal{S}$ to generate h_G. But as observed in Taskonomy [67], it generally suffers from lack of generalization.

Recent approaches such as Taskonomy [67, 2] have addressed this question by selecting the $\mathcal{K}_{(t_i,d_j)}$ where t_i and t_j are most correlated.

In our approach (see Section 3.2), $\mathcal{G}$ is a neuronal encoder allowing to estimate a representation h_G of reduced dimensionality, still approximating well the h_i and potentially leading to better generalizations by removing biases due to over-complexity.

Nevertheless, $\mathcal{E}^G$ is then composed of all $\mathcal{E}_{(t^S_i,d^S_i)}$ and of $\mathcal{G}$ and thus mapping a given x to h_G has a high computational cost. In Section 3.3 we propose the use of *distillation* to transform $\mathcal{E}^G$ into a unique and lightweight model $\mathcal{E}^G_{unique}$, directly projecting x to $\hat{h}_G$ and therefore allowing easier transfer when using it to get a new knowledge.

Source and target knowledge for facial analysis As building a general knowledge on all tasks and domains existing would not be feasible, we propose to validate our approach experimentally by carefully designing a set $\mathcal{S}$ of Source-Knowledge and a set $\mathcal{T}$ of Target-Knowledge to cover a facial analysis general knowledge, including the 15 different knowledge described in Table 1.

$\mathcal{S}$ gathers M=6 source knowledge, which leads to a domain $\bigcup_i d^S_i$ of around 22 million faces with M=6 different tasks contains in around 106 million parameters. To train $\mathcal{G}$ (in an unsupervised fashion), instead of using $\bigcup_i d^S_i$, we will use a domain $\mathcal{D}_{unsup}$ of 4.12 millions of faces (discarding annotations), extracted from VGGFace2 [10] (3.14 million), EmotioNet [24] (0.72 million) and IMDb-WIKI [57] (0.26 million).

$\mathcal{T}$ contains 9 various knowledge on facial analysis. Note that two tasks with the same name (for instance Expression Classification) may still be different (for instance facial expression is subjective and depends on its annotators [68]).

$\mathcal{K}_{(t_i,d_i)}$	t_i	d_i size
Expr-AffectNet	Emotion Classif. [63]	0.3M [47]
Expr-RAF	Emotion Classif. [63]	15,339 [42]
Expr-SFEW	Emotion Classif. [63]	1,766 [22]
Identity-MS	Identity Matching [60]	6.5M [28]
Identity-LFW	Identity Matching [60]	13,000 [33]
Gender-IMDb	Gender Prediction [4]	0.5M [57]
Gender-UTK	Gender Prediction [20]	20,000 [44]
Attrib-CelebA	Attrib. Detection [9]	0.2M [44]
AgeR-IMDb	Age Regression [57]	0.5M [57]
AgeR-FG	Age Regression [4]	1,000 [38]
AgeR-UTK	Age Regression [20]	20,000 [71]
AgeC-UTK	Age Classif. [20]	20,000 [71]
Ethnic-UTK	Ethnicity Classif. [20]	20,000 [71]
Pain-UNBC	Pain Estimation [72]	48,000 [45]
Object-ImageNet	Object Classif. [21]	14M [21]

Table 1. The M=6 source and 9 target knowledge (source is in bold). The task, as well as the used pre-trained model for source knowledge, are described in the associated papers. Information about the used domain can be found in the second citation on each row. Moreover, more details about all the selected knowledge will be provided in supplementary materials.

3.2. Dimensionality Reduction for a More General Knowledge

We explain in the previous subsection that we dispose of M source knowledge encoders $\mathcal{E}_{(t^S_i,d^S_i)}$, allowing to extract M h_i embedding from a given face x. We study here

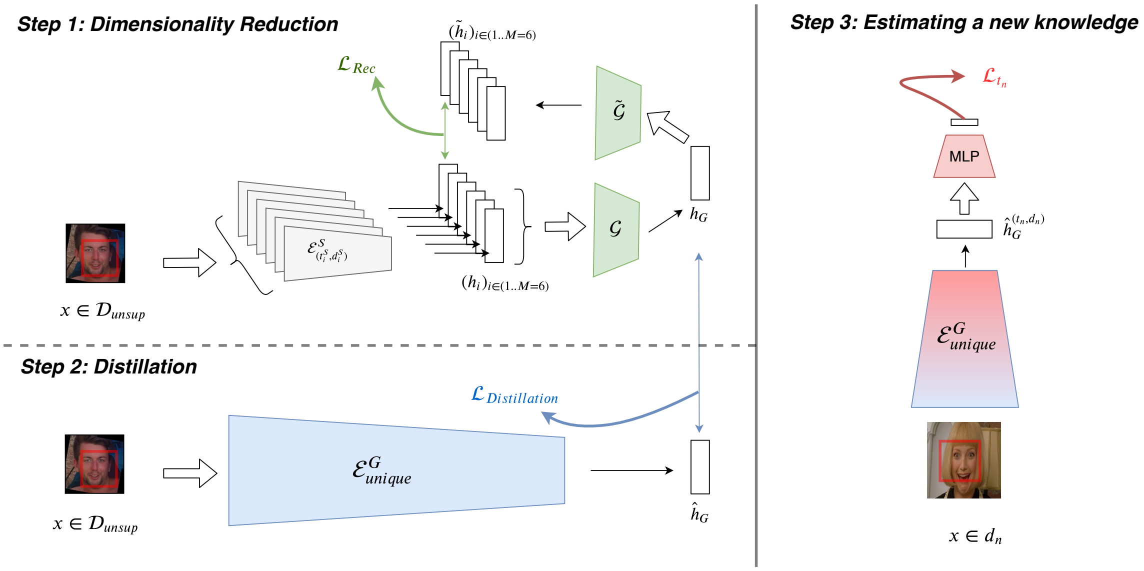

Figure 1. Overview of the method. *Step 1* is the training of $\mathcal{G}$ on $\mathcal{D}_{unsup}$, to obtain an encoder $\mathcal{E}^G$ of x in a compact embedding h_G, gathering all sources of knowledge (represented by the extracted h_j). *Step 2* aims at reducing the number of parameters used to get h_G by distilling the knowledge of 130 millions parameters into a much simpler encoder $\mathcal{E}^G_{unique}$, leading to $\hat{h}_G$. In *step 3* this lightweight encoder $\mathcal{E}^G_{unique}$ can then be plugged with a MLP $\mathcal{C}_{(t_n,d_n)}$ and all parameters adapted to learn new knowledge $\mathcal{K}_{(t_n,d_n)}$.

the operator $\mathcal{G}$ allowing to combine these embedding into a compact and general embedding h_G.

Motivations Defining $\mathcal{G}$ as a basic concatenation function implies a very large and redundant h_G. Therefore there is a meaning in reducing the dimension of h_G, leading to discard some redundancies and exploit the complementarity between the h_j. Dimensionality reduction is a well-explored topic, from linear projections such as PCA [35] to more complex non-linear approaches such as manifold learning [34]. The major drawback of a linear approach is that the representation is projected into a plane and therefore can miss the real shape of the data. Thus, learning the manifold represented by h_G with a (non-linear) neural network [41] makes sense.

Adopted Approach As we want to reduce the dimensionality without supervision, we propose to train $\mathcal{G}$ with an auto-encoding objective $\mathcal{L}_{Rec}$, as shown in Step 1 of Figure 1. In other words, we are building h_G as an answer to all source tasks (represented by the h_j) but on the $\mathcal{D}_{unsup}$ domain. For that, we optimize the parameters of both $\mathcal{G}$ (as an encoder) and of $\tilde{\mathcal{G}}$ (as a decoder reconstructing the h_i):

$$h_G = \mathcal{G}((h_i)_{i=1..M}) \tag{1}$$

$$(\tilde{h}_i)_{i=1..M} = \tilde{\mathcal{G}}(h_G) \tag{2}$$

$$\mathcal{L}_{Rec} = \Sigma_{m=1}^{M} ||\hat{h}_i - h_i||^2 \tag{3}$$

Note that we choose $\tilde{\mathcal{G}}$ to have a symmetric architecture to $\mathcal{G}$. We will further discuss the architecture choice of an autoencoder in the experiments section, by also experimenting with PCA, regular autoencoders [7], variational autoencoders [37] and denoising autoencoders [64].

Another important choice is the choice of the dimensionality of h_G, as it drives the knowledge compression process. As we are aiming at creating a unique general knowledge, we consider as a rule of thumb to fix it as the average dimensionality of the h_i. We will conduct a empirical study in subsection 4.4 on the impact of this dimension on the quality of h_G and check that there is no optimal dimension, only extreme values (very low-dimensional or high-dimensional) clearly degrading our approach.

3.3. Real-world Transfer Learning by Distillation

As mentioned in the problem definition, obtaining h_G by step 1 implies using $\mathcal{E}^G$, composed of M=6 different pre-trained models combined to $\mathcal{G}$. It leads to a huge number of parameters (130M) and greatly limits the possibility of domain adaptation when dealing with the target knowledge, as we can't adapt such a large number of parameters on a new domain. A natural way to solve this problem would then to compress the so-obtained model. Model compression can be addressed with several approaches as shown in the recent literature [18], mainly gathered into four categories: parameters pruning, low-rank factorization, compact convolutional filters and knowledge distillation.

Our current model is composed of M=6 different specific branches and we do not only want to reduce the number of parameters: we want to achieve a unique encoder. It is the promise of the distillation approach, allowing to train a new *Student* model, supervised by the previous big *Teacher* model [32]. Thus, we will not only make $\mathcal{E}^G$ lighter but also transform it to a conventional estimator $\mathcal{E}^G_{unique}$, such as a classic convolutional neural network architecture, on which well-known methods such as data augmentation can be easily applied. Moreover previous works [19] have shown that fine-tuning all the parameters of a model may lead to better domain adaptation and improve the quality of the transfer.

Distillation To achieve distillation, we still are training our new $\mathcal{E}^G_{unique}$ on $\mathcal{D}_{unsup}$, as illustrated in Figure 1. We consider $\mathcal{E}^G_{unique}$ as a neural network of arbitrary architecture (we choose a ResNet-18 [31] for all experiments), taking the face image x and directly projecting it into a representation $\hat{h}_G$.

The training is down by minimizing $\mathcal{L}_{Distillation}$

$$\mathcal{L}_{Distillation} = D_c(h_G, \hat{h}_G) \qquad (4)$$

where D_c is the cosine metric (*i.e.* $D_c(a,b) = \frac{a^\top b}{\|a\| \cdot \|b\|}$).

Using $\mathcal{E}^G_{unique}$ to estimate knowledge from $\mathcal{T}$ Finally when $\mathcal{E}^G_{unique}$ has been trained, the last step is straight-forward and only consists in adding a Multi-Layer-Perceptron on top of it and train all parameters on to estimate the new target knowledge.

3.4. About Using 2 Step Training

The two previous steps may be seen as training two parts of the same model and may be done at the same time, by training $\mathcal{E}^G_{MT} = \tilde{\mathcal{G}} \circ \mathcal{E}^G_{unique}$ to directly fit the h_i. This approach may be considered as a **weakly supervised multi-task (MT) learning**. In practice we observe that it does not converge as well as our two-step approach, which is disentangling the processing relative to the tasks (first step) and to the domains (second step). It also is in line with several progressive approaches observed in the literature [36], achieving better model convergence by progressively training different parts of the model.

4. Experiments

This section validates our approach, first by detailing results of $\mathcal{E}^G_{unique}$ on the different source and target knowledge, compared to state-of-the-art dedicated approaches, then by running an ablation study on the different steps of our method and on the benefit to decompose the learning process into these steps.

Domain	# Detected Faces	# Undetected Faces
CelebA	202442	177
UTKFace	24018	98
FG-NET	1002	0
SFEW	1732	37
RAF	15330	9
ShoulderPain	48391	0
LFW	13233	0
$\mathcal{D}_{unsup}$	4.12 M	0.02M

Table 2. Number of detected/undetected faces for each domain.

4.1. Implementation details

We pre-process all faces from all domains with same operations. We use a private face detector to first detect and loosely crop the detected faces. If more than one face is detected, we select the closest to the image center. A landmark-based aligner is then applied on the detected faces, which are finally resized to 300 × 300 × 3 pixels. Table 2 reports the number of images where a face was detected for each domain (first column). When a face is not detected (second column), we will consider during evaluation of our models (a) that the prediction is not correct when addressing classification problems (b) and that the prediction is the average between minimum and maximum values when dealing with regression problems. Note that the rest of the paper is following this rule.

As we use different tasks and domains, the evaluation protocols are changing for each knowledge. We follow the exact evaluation protocol used by the state-of-the-art dedicated approaches we are comparing to.

4.2. Experimental Validation

We first propose to evaluate our final approach $\hat{h}^{(t,d)}_G$ on target knowledge. Looking at the two last columns of Table 3, we observe that our approach adapt well to all target knowledge, even outperforming dedicated state-of-the-art approaches on AgeC-UTK, Expr-SFEW, Expr-RAF, Ethnic-UTK, AgeR-UTK. Note that the model for $\hat{h}^{(t,d)}_G$ counts only 2.2 millions parameters, which is almost always far less than the parameters used by other approaches, helping to better generalize and better fits real-world applications constraints.

Target Knowledge Moreover, this level of performance with a lightweight model is difficult to achieve, as shown by the results of the CNN baseline, which has the same architecture than $\hat{\mathcal{K}}^G_{unique}$ but is trained from scratch on the domain of the target knowledge. Indeed, this CNN is not always converging (*e.g.* on AgeR-FG with only 1002 images) and there is a huge gap between its scores and our approach, highlighting the clear benefit of transferring knowledge. A second stronger baseline called P-CNN is basically the clas-

Dataset	Metric	CNN	P-CNN	$\mathcal{E}^G$	$\mathcal{E}^G_{MT}$		$\mathcal{E}^G_{unique}$		State-of-the-art (# parameters)
				h_G	$\hat{h}_{MT}$	$\hat{h}^{(t,d)}_{MT}$	$\hat{h}_G$	$\hat{h}^{(t,d)}_G$	
AgeC-UTK	Acc.	57.80	64.30	68.80	67.90	68.40	68.80	**70.40**	70.10 [20](5 M)
Gender-UTK	Acc.	90.00	96.50	97.20	93.15	94.4	97.58	97.90	**98.23 [20]** (5 M)
Expr-SFEW	Acc.	22.00	53.60	52.10	50.20	52.20	54.00	**57.2**	**55.40-58.14 [63] [1]** (1.7 M/5 M)
Expr-RAF	Acc.	69.00	85.70	86.51	82.40	87.20	87.30	**89.3**	86.77 [68] (35 M)
Ethnic-UTK	Acc.	62.20	84.90	89.20	81.20	82.20	88.2	**91.20**	90.10 [20] (5 M)
Identity-LFW	Acc.	(98.40)	(99.65)	99.10	94.27	(99.1)	98.92	(99.42)	**99.65-99.87** [60](25 M)
AgeR-UTK	MAE	6.38	4.70	4.39	4.70	4.42	4.24	**4.05**	5.39 [11](21.8 M)
Pain-UNBC	MAE	0.89	0.56	0.54	0.64	0.72	0.56	0.52	**0.51 [72]** (0.0001 M)
AgeR-FG	MAE	11.10	3.10	**2.85**	3.95	3.12	2.95	3.05	**2.81-3.00** [58] [4] (25 M)
Attrib-CelebA	ER	8.60	8.04	7.70	8.12	8.04	7.81	7.67	**7.02 [9]**(16 M)

Table 3. Performance of diverse end-to-end approach evaluated on the target knowledge. All approaches are described in the methods or detailed in the text. Note that for the specific case of LFW, two types of results are reported. For result between parenthesis, before evaluation on LFW, the network is first trained (or fine-tuned) on a subset of 100,000 faces from VGGFace2 to predict Identity. The other LFW results are obtained directly from embedding h. Acc. is Accuracy and ER is the Average Error Rate.

sic way for transfer learning: it consists in the use of a pre-trained CNN, chosen among the source encoders as the one transferring the best for the target task and then fine-tuned on the target task for all its parameters. See also Table 6 for the scores obtained by selecting the best source encoder but without fine-tuning all parameters. Note that fine-tuning all parameters does not lead always to an improvement.

Source Knowledge We also analyze the performance of $\hat{h}^{(t,d)}_G$ on the source knowledge and compare it to the original pre-trained models used as source knowledge estimators in Table 4. For Expression-AffectNet, Attrib-CelebA and Gender-UTK we observe a clear improvement, while for Age-FG and for Identity-LFW the performance is slightly degraded. For Age-FG it may be explained by performance saturation, as the domain is very small and the human performance is around 4.6 in MAE [29]. For Identity-LFW, the loss can come from the difference in representation size: the original model representation has 2048 features dedicated to Identity Matching, while our model counts only 1024 features. Augmenting the dimension of $\hat{h}_g$ may lead to more comparable performance and will be discussed in subsection 4.4. Finally, to avoid a long training time we do not compare the models on ImageNet but on the smaller dataset TinyImageNet [66]. The obtained accuracy (on the validation set) clearly underlines the importance in the choice of $\mathcal{D}_{unsup}$: the unsupervised training has been done only on face images and thus is not beneficial for tasks such as ImageNet classification. Despite this loss in performance, note that when training a ResNet-18 from scratch on TinyImageNet, we achieve only 52% accuracy.

4.3. Ablation study

Contribution of $\mathcal{G}$ As described in Section 3, $\mathcal{G}$ takes as input the concatenation of the 6 embeddings (h_i), one per task (of size 5488 in our case). Then, the goal is to generate a compact representation h_G (which size is fixed to 1024 in all the experiments) of the M=6 embeddings (h_i). As this dimension reduction step is a crucial operation of our approach, we propose to study the impact of different variations on this step.

Knowledge	Metric	Original	$\mathcal{E}^G$	$\mathcal{E}^G_{unique}$
Expr-AffectNet	Accuracy	63.5	64.0	**64.4**
Identity-LFW	Accuracy	**99.65**	99.1	99.42
Gender-UTK	Accuracy	96.5	97.2	**97.9**
Object-TinyImageNet	Accuracy	**76.2**	71.8	56.5
AgeR-FG	MAE	**2.85**	**2.85**	3.05
Attrib-CelebA	Error rate	8.03	7.7	**7.67**

Table 4. Performance of the original models $\mathcal{E}^S_{(t^S_j,d^S_j)}$, of $\mathcal{E}^G$ and of $\mathcal{E}^G_{unique}$ on the original tasks.

Therefore, we first evaluate the ability of $\mathcal{G}$ to reconstruct the original h_i features from h_G. We report in Table 5 the normalized Root Mean Squared Error between each h_i and $\tilde{h}_i$ with diverse variations on $\mathcal{G}$. PCA stands for Principal Components Analysis, AE for standard Autoencoder, VAE for Variational Autoencoder and DAE for Denoising Autoencoder. Note that AE, VAE and DAE have the same number of parameters (the encoder $\mathcal{G}$ has 3 fully connected layers: 5488×3136, 3136×1792 and 1792×1024 and the decoder $\tilde{\mathcal{G}}$ is the symmetrical). The linear PCA baseline is easily outperformed by other approaches. The two best performing are the AE and the DAE. Surprisingly, the DAE does not generalize better than the AE, and is performing better only for the reconstruction of the ImageNet embedding, which is the less relevant in a context of face inputs. Still note that the gap of performance is not very important and all non-linear methods allows a fairly decent reconstruction of the embeddings.

h_i	PCA	AE	VAE	DAE	$\mathcal{E}^G_{MT}$	$\mathcal{E}^G_{unique}$
Expr	0.28	**0.23**	0.26	0.26	0.55	0.39
Identity	0.25	**0.22**	0.23	0.26	1.12	0.35
Object	0.48	0.26	0.30	**0.25**	0.67	0.44
Age	0.23	**0.19**	0.22	**0.19**	0.77	0.38
Attrib	0.33	**0.27**	0.31	0.29	0.97	0.35
Gender	0.25	**0.25**	0.27	0.27	1.22	0.30
Average	0.31	**0.24**	0.27	0.25	0.92	0.37

Table 5. Reconstruction normalized RMSE obtained on $\mathcal{D}_{unsup}$ test set by considered methods for each source knowledge embedding h_i and on average.

Contribution of the distillation: $\mathcal{E}^G_{MT}$ versus $\mathcal{E}^G_{unique}$ The last two columns of the Table 5 allow to evaluate $\mathcal{E}^G_{MT}$ and $\mathcal{E}^G_{unique}$, by their ability to reconstruct the h_i from the embedding they produced. To achieve such a reconstruction, we extracted the $\hat{h}_{MT}$ and $\hat{h}_G$ from all element of D_{unsup} and then trained a decoder (with a similar architecture to $\tilde{\mathcal{G}}$) to reconstruct the h_i.

We can observe that the reconstruction error of $\mathcal{E}^G_{MT}$ is sometimes very high (*e.g.* for the identity embedding, explaining the low results of $\hat{h}_{MT}$ when use for Identity-LFW in Table 3). Note that simply generating random embeddings (according to a uniform distribution in the range of the target embeddings) allows to achieve a normalized RMSE of 1.4. Therefore a score of 1.22 is almost random.

Without surprise, the reconstruction error of $\mathcal{E}^G_{unique}$ is higher than the one obtained by $\mathcal{G}$, which can be explained by both the distillation approximation between h_G and $\hat{h}_G$ and the lower capacity of our small model. Nevertheless, a low reconstruction error is not a guarantee of better performance when trying to apply the h_G representation in other tasks and domains. Thus, we propose to evaluate the very transfer learning operation using $\mathcal{E}^G$ instead of $\mathcal{E}^G_{unique}$. The first column of Table 3 reports the results obtained by $\mathcal{E}^G$ on the target knowledge, while Table 4 provides the results on the source knowledge. On almost all knowledge, $\mathcal{E}^G$ is outperformed by $\mathcal{E}^G_{unique}$ by a significant margin.

Yet, if we try to understand from where the distillation improvement is coming, we can observe that only using a MLP on top of $\hat{h}_G$ (without fine-tuning all parameters) does not bring improvements compared with using a similar MLP on top of h_G, some small changes in term of performance being observed. Thus, the distillation error has a limited impact when dealing with transfer. The bigger improvement is observed when using $\hat{h}^{(t,d)}_G$, meaning that we fine-tune all parameters of $\mathcal{E}^G_{unique}$ on the new knowledge. It illustrates that the main contribution of the distillation lies in the reduced size of the encoder, giving it the ability to more easily adapt to new tasks and domains.

Knowledge	$\mathcal{G}$			Selection	
	Concat	PCA	AE	BT	BCT
AgeC-UTK	65.50	65.00	**68.80**	63.2	67.20
Gender-UTK	96.92	96.7	**97.20**	96.50	97.10
Expr-SFEW	45.70	32	52.10	52.20	**53.1**
Expr-RAF	84.89	81.9	**86.51**	85.48	85.74
Ethnic-UTK	86.65	62.5	**89.20**	83.40	86.20
Identity-LFW	88.10	96.8	99.10	99.65	**99.7**
AgeR-UTK	4.45	4.68	**4.39**	4.70	4.54
Pain-UNBC	0.69	0.6	0.54	0.53	**0.51**
AgeR-FG	3.22	3.18	**2.85**	**2.85**	**2.85**
Attrib-CelebA	8.07	8.03	**7.70**	8.03	7.85

Table 6. Performance of different variations of $\mathcal{G}$ and of selection approaches on the 9 targets knowledge (performance metrics are the same used in Table 3).

Concatenation, reduction with $\mathcal{G}$ or selection ? If we come back to the operator $\mathcal{G}$, we discussed in section 3 the good reasons to reduce dimensionality of the h_i. We propose to empirically compare this choice to other methods and validate the intuition that selection may discard useful information. Thus, for each target knowledge, we propose to evaluate $\mathcal{G}$ and several alternatives summarized in Table 6. We first evaluate a Multi Layer Perceptron taking as input the concatenation (Concat of all h_i) and having $\mathcal{G}$ architecture. The MLP parameters are trained from scratch on the target knowledge. Our $\mathcal{G}$ is then evaluated by training a small MLP on top of h_G (extracted from all images of the target domain). We also compare the benefits brought by using a non-linear $\mathcal{G}$ by also reporting the Princpal Components Analysis (PCA) results.

Finally, another concurrent approach consists in selecting the embedding (resp. the combination of embedding) yielding the Best Transfer (BT) (resp. the Best Combination Transfer (BCT)). Several works [67, 40] proposed a method to automatically select such a best combination of embeddings. We choose here to reproduce this approach in a naive way, by brute force testing all the possible single transfers (BT) or combinations (by concatenation) of transfers (BCT) and reporting the best found results on each target kowledge in the two last columns of Table 6.

Table 6 shows that the Concat baseline is outperformed by all other approaches on almost all target knowledge. It is in line with what Zamir *et al.* [67] observed and it might be explained by the large dimensionality of the input of the Concat method and the limited size of some of the target domains. Then, it is interesting to see that the simple PCA method is nevertheless several times on par with the BT results, while $\mathcal{G}$ performances are better than BT results and often better than the BCT. Thus, it validates the choice of a reduction of dimensionality applied on all embedding h_i, moreover illustrating the intuition developed in introduc-

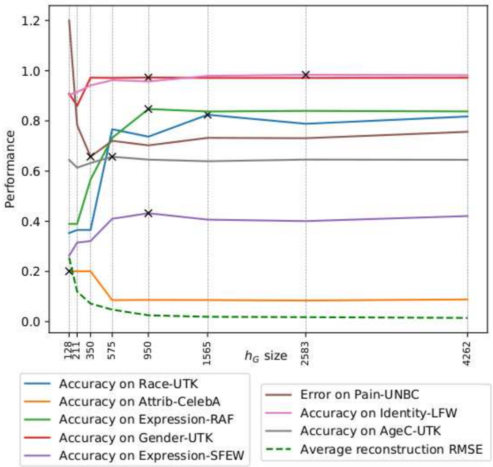

Figure 2. Impact of the size of h_G on the reconstruction error of $\tilde{\mathcal{G}}$ and on the transfer for some of the target knowledge. Black cross is the optimal representation size for a given knowledge.

tion: *when the embeddings from which we want to transfer knowledge are correlated, there is a benefit to exploit these redundancies instead of discarding some information by block*, as done by the selection methods. Note that in contrast to the direct Concat baseline approach, $\mathcal{G}$ is able to extract a meaningful h_G from the h_i because of the large number of observed samples of $\mathcal{D}_{unsup}$.

4.4. Impact of h_G dimension

During the presentation of the method, we propose to choose the size of h_G as the average size of h_i. In Figure 2 we discuss this choice, by showing that there is no optimal representation size for all knowledge. Nevertheless, a too small representation conduct to low results and a large representation implies a high-computational cost and lower performance.

4.5. Effect of Learning in 2 Stages

During the construction of our model, we argue that there was a benefit to adopt a 2 stage approach, disentangling the compression and the distillation steps. We already have observed the low ability of h_{MT} to reconstruct the source embedding h_i. Yet, what is the real impact in term of transfer learning of this insufficient convergence.

Let's study in details the columns dedicated to $\mathcal{E}^G_{MT}$ and $\mathcal{E}^G_{unique}$ in Table 3. We observe that if we are not fine-tuning the whole parameters of the model and only using the h_{MT} and $\hat{h}_G$ as features, the gap is significant on most of the source and target knowledge. Moreover, these results are correlated task by task to the reconstruction error observed on each h_i. Nevertheless h_{MT} allows to achieve far better performance than the CNN baseline, still underlining the benefit of a pre-training, even when it is a noisy one.

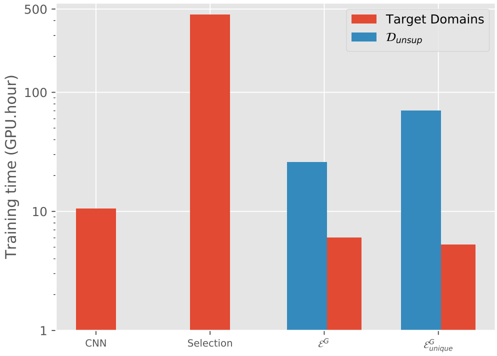

Figure 3. Total training time in hours (logscale) with one P100 GPU for different methods, on $\mathcal{D}_{unsup}$ and on all target domains.

4.6. Training Time

Finally, we report from Figure 3 the training time spent on $\mathcal{D}_{unsup}$ and on the tartget domains for different approaches. Even if the training on $\mathcal{D}_{unsup}$ is relatively long (almost 70 hours), the transfer learning step is then really faster than training a model from scratch. Moreover compared to the brute-force baseline of selecting all possible combinations (BCT), the total training time of the $\mathcal{E}^G_{unique}$ (taking the training time of $\mathcal{G}$ into account) is divided by 4. We can also project ourselves in the case where we had much more knowledge to master. For instance, multiplying the number of target knowledge by 2 will imply a factor of time of 8 between $\mathcal{E}^G_{unique}$ and BCT.

5. Conclusions

This paper has introduced a novel approach to the problem of multi-source transfer learning, validated in the context of facial analysis. A unique and general model $\mathcal{E}^G_{unique}$, obtained by merging six different source knowledge, can be transferred on 9 different target knowledge. Building this model is done through two successive training steps. First, an autoencoder $\mathcal{G}$ if trained to combine the hidden representations of the existing models into one single unifying embedding h_G. Then, distilling this model to a light-weight student CNN allows to reduce the number of parameters and improve the adaptation ability of the model. The approach was experimentally validated by an exhaustive ablation study and performances on par with state-of-the-art methods on the 15 different knowledge, with a single simple model. On overall, the approach provides an efficient way to obtain universal models compressing the knowledge included in several existing models, without loss in performance, allowing an easy exploitation in real-world applications.

References

[1] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool. Covariance pooling for facial expression recognition. In *CVPR Workshop*, 2018. 6

[2] A. Achille, M. Lam, R. Tewari, A. Ravichandran, S. Maji, C. Fowlkes, S. Soatto, and P. Perona. Task2vec: Task embedding for meta-learning. *arXiv preprint arXiv:1902.03545*, 2019. 1, 3

[3] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *ICML*, 2013. 2

[4] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay. Effective training of convolutional neural networks for face-based gender and age prediction. *Pattern Recognition*, 2017. 3, 6

[5] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 2010. 2

[6] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *NIPS*, 2016. 2

[7] D. H. Ballard. Modular learning in neural networks. In *AAAI*, 1987. 4

[8] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *T-PAMI*, 2019. 2

[9] J. Cao, Y. Li, and Z. Zhang. Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In *CVPR*, 2018. 3, 6

[10] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018. 3

[11] W. Cao, V. Mirjalili, and S. Raschka. Consistent rank logits for ordinal regression with convolutional neural networks. *arXiv preprint arXiv:1901.07884*, 2019. 6

[12] Z. Cao, M. Long, J. Wang, and M. I. Jordan. Partial transfer learning with selective adversarial networks. In *CVPR*, 2018. 2

[13] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018. 2

[14] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran. Correlational neural networks. *Neural computation*, 2016. 2

[15] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni. Deep, landmark-free fame: Face alignment, modeling, and expression estimation. *IJCV*, 2019. 2

[16] Y. Chebotar and A. Waters. Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech*, pages 3439–3443, 2016. 2

[17] S. Chen, C. Zhang, and M. Dong. Coupled end-to-end transfer learning with generalized fisher information. In *CVPR*, 2018. 2

[18] Y. Cheng, D. Wang, P. Zhou, and T. Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017. 4

[19] B. Chu, V. Madhavan, O. Beijbom, J. Hoffman, and T. Darrell. Best practices for fine-tuning visual classifiers to new domains. In *ECCV Workshop*, 2016. 5

[20] A. Das, A. Dantcheva, and F. Bremond. Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach. In *ECCV Workshop*, 2018. 3, 6

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[22] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *ICCV Workshop*, 2011. 3

[23] C. Doersch and A. Zisserman. Multi-task self-supervised visual learning. In *ICCV*, 2017. 2

[24] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, 2016. 3

[25] R. C. Geyer, V. Wegmayr, and L. Corinzia. Transfer learning by adaptive merging of multiple models. In *MIDL*, 2019. 2

[26] M. Ghifary, W. Bastiaan Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, 2015. 2

[27] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 2

[28] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 3

[29] H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In *ICB*, 2013. 6

[30] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004. 2

[31] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[32] G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 5

[33] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 3

[34] X. Huo, X. S. Ni, and A. K. Smith. A survey of manifold-based learning methods. *Recent advances in data mining of enterprise data*, 2007. 4

[35] I. Joliffe and B. Morgan. Principal component analysis and exploratory factor analysis. *Statistical methods in medical research*, 1(1):69–95, 1992. 2, 4

[36] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 5

[37] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 4

[38] A. Lanitis and T. Cootes. Fg-net aging data base. *Cyprus College*, 2(3):5, 2002. 3

[39] S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang. Overcoming catastrophic forgetting by incremental moment matching. In *NIPS*, 2017. 2

[40] J. Li, P. Zhou, Y. Chen, J. Zhao, S. Roy, Y. Shuicheng, J. Feng, and T. Sim. Task relation networks. In *WACV*, 2019. 1, 7

[41] S. Li. Measure, manifold, learning, and optimization: A theory of neural networks. *arXiv preprint arXiv:1811.12783*, 2018. 4

[42] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, 2017. 3

[43] X. Li, H. Xiong, H. Wang, Y. Rao, L. Liu, and J. Huan. Delta: Deep learning transfer using feature map with attention for convolutional networks. In *ICLR*, 2019. 2

[44] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 3

[45] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *FG*, 2011. 3

[46] M. Mancini, S. R. Bulò, B. Caputo, and E. Ricci. Best sources forward: domain generalization through source-specific nets. In *ICIP*, 2018. 2

[47] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *Transactions on Affective Computing*, 2017. 3

[48] N. Neverova, C. Wolf, G. Taylor, and F. Nebout. Moddrop: adaptive multi-modal gesture recognition. *T-PAMI*, 2016. 2

[49] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011. 2

[50] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 2

[51] M. Noroozi, H. Pirsiavash, and P. Favaro. Representation learning by learning to count. In *ICCV*, 2017. 2

[52] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *ICLR*, 2017. 2

[53] J.-M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie. Mfas: Multimodal fusion architecture search. In *CVPR*, 2019. 2

[54] F. Radenovic, G. Tolias, and O. Chum. Deep shape matching. In *ECCV*, 2018. 2

[55] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He. Data distillation: Towards omni-supervised learning. In *CVPR*, 2018. 2

[56] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015. 2

[57] R. Rothe, R. Timofte, and L. Van Gool. Dex: Deep expectation of apparent age from a single image. In *ICCV Workshop*, 2015. 3

[58] R. Rothe, R. Timofte, and L. Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 2018. 6

[59] S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard. Latent multi-task architecture learning. In *AAAI*, 2019. 2

[60] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 3, 6

[61] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang. Deep multimodal feature analysis for action recognition in rgb+ d videos. *T-PAMI*, 2018. 2

[62] C. G. Snoek, M. Worring, and A. W. Smeulders. Early versus late fusion in semantic video analysis. In *ACMMM*, 2005. 2

[63] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy, and F. Jurie. An occam's razor view on learning audiovisual emotion recognition with small training sets. In *ICMI*, 2018. 3, 6

[64] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 2010. 4

[65] D. Xu, W. Ouyang, X. Wang, and N. Sebe. Pad-net: multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, 2018. 2

[66] L. Yao and J. Miller. Tiny imagenet classification with convolutional neural networks. *CS 231N*, 2015. 6

[67] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 1, 2, 3, 7

[68] J. Zeng, S. Shan, and X. Chen. Facial expression recognition with inconsistently annotated datasets. In *ECCV*, 2018. 3, 6

[69] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *ECCV*, 2016. 2

[70] R. Zhang, P. Isola, and A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017. 2

[71] S.-Y. Zhang, Zhifei and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 2017. 3

[72] Y. Zhang, R. Zhao, W. Dong, B.-G. Hu, and Q. Ji. Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation. In *CVPR*, 2018. 3, 6