人脸识别技术介绍和表情识别最新研究

一、人脸识别技术介绍

人脸识别作为一种生物特征识别技术，具有非侵扰性、非接触性、友好性和便捷性等优点。早在二十世纪初期，人脸识别已经出现，于二十世纪中期，发展成为独立的学科。人脸识别真正进入应用阶段是在 90 年代后期。人脸识别属于人脸匹配的领域，人脸匹配的方法主要包括特征表示和相似性度量。

人脸识别通用的流程主要包括人脸检测、人脸裁剪、人脸校正、特征提取和人脸识别。人脸检测是从获取的图像中去除干扰，提取人脸信息，获取人脸图像位置，检测的成功率主要受图像质量，光线强弱和遮挡等因素影响。获取人脸后，人脸裁剪是根据实际需求，裁剪部分或整体的人脸，进一步精确化人脸图像。为提高人脸识别准确率，人脸校正可以尽可能的降低由于姿态和表情导致的人脸变化，获取正面或者平静状态下的人脸照片。特征提取利用不同的特征，对图片进行相似度的衡量和评价。人脸识别主要包括一对一或者一对多的应用场景，对目标人脸进行识别和验证。

人脸表达模型主要分为 2D，2.5D，3D。2D 人脸指的是 RGB，灰度和红外图像，是确定视角下表征颜色或纹理的图像，不包括深度信息。2.5D 是在某一视角下拍摄获取的人脸深度数据，但是曲面信息不连续，没有被遮挡部分的深度数据信息。3D 人脸由多张不同角度的深度图像合成，具有完整连续的曲面信息，包含深度信息。2D 图像人脸识别的研究时间较长，软硬件技术较为完备，得到了广泛的应用。但是由于 2D 图像反映二维平面信息，不包含深度数据，不能够完整的表达出真实人脸模型。相比于二维人脸图像，三维图像不受光照等影响，具有更强的描述能力，能够更为真实的反映人脸信息，在人脸合成、人脸迁移、三维人脸识别等场景中应用。3D 人脸识别一般采用深度相机获取人脸深度信息，主要包括双目相机，基于结构光原理的 RGB-D 相机和基于光飞行时间原理的 TOF 相机。常见的三维人脸识别算法主要包括传统识别方法和深度学习识别方法。

1.传统识别方法

(1)基于点云数据的人脸识别

点云是 3D 人脸数据的一种表征方式，每一个点都对应一个三维坐标，扫描设备使用这种数据格式存储采集的三维人脸信息，甚至可以将稀疏坐标也拼接到形状信息上，更为完善的反映人脸信息。基于点云数据的 3D 人脸识别直接使用三维点云进行匹配，常见方法有 ICP(Iterative Closest Point)和 Hausdorff 距离。前者可以修正点云信息中平移和旋转变换的误差，后者利用三维点云之间的距离最大值，匹配人脸，但是两者均存在鲁棒性不足的问题。

(2)基于面部特征的 3D 人脸识别

人脸的面部特征主要包括局部特征和全局特征，局部特征可以选择从深度图像上提取关于面部关键点的特征信息，全局特征是对整张人脸进行变换提取特征，例如球面谐波特征或者稀疏系数特征。

2.深度学习识别方法

(1)基于深度图的人脸识别

深度图像中三维数据的 z 值被投影至二维平面，形成平滑的三维曲面。可使用归一化网络和特征提取网络实现深度图人脸识别，归一化网络将输入的深度图像转化为 HHA 图像，再使用卷积神经网络回归用于获取归一化深度图的参数，特征提取网络用于获取表征深度图人脸的特征向量。

(2)基于 RGB-3DMM 的人脸识别

3DMM 是指三维人脸变形统计模型，其最早是用于解决从二维人脸图像恢复三维形状

的问题，现多被用于对深度图像或彩色图像进行人脸模型回归，实现识别任务。

(3)基于 RGB-D 的人脸识别

RGB-D 图像是包含了彩色图像和深度图，前者是从红、绿、蓝颜色通道获取的图像，后者是指包含与视点的场景对象的表面的距离有关的图像通道，两者之间是相互配准。通过对彩色图像和多帧融合后的深度图像分别进行预训练和迁移学习，在特征层进行融合，提高人脸识别率。

二、表情识别最新研究

1) Facial Emotion Recognition with Noisy Multi-task Annotations

摘要

从面部表情可以推断出人类的情感。但是，在常见的情感编码模型中，包括分类和维度模型，面部表情的注释通常会非常嘈杂。为了减少人为标注多任务标签的工作量，文中引入了带有嘈杂的多任务注释的面部表情识别新问题。对于这个新问题，文中建议从联合分布匹配的角度进行计算，其目的是学习原始人脸图像和多任务标签之间更可靠的关联，从而减少噪声影响。采用一种新方法来在统一的对抗性学习游戏中启用情绪预测和联合分布学习。在广泛的实验中进行的评估研究了所提出的新问题的实际设置，以及所提出的方法在合成嘈杂的带标签CIFAR-10或实际嘈杂的多点干扰方法上优于最新竞争方法的明显优势标记为RAF和 AffectNet 的任务。

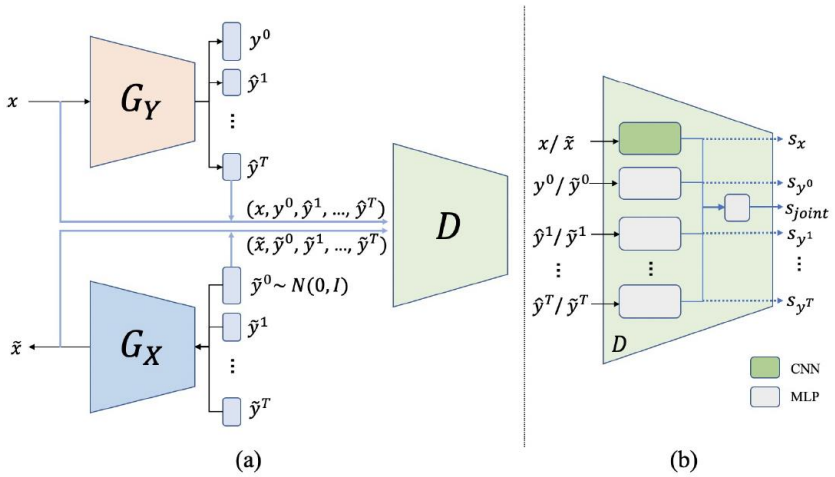
本文探讨的是嘈杂的多任务标签中面部表情识别的问题。实际应用中，两种最常用的面部情绪编码模型是分类和维数，但是通过从可用的情感标签中进行模型的学习容易产生不好的结果，因此，文中提出的公式是从联合分布匹配的角度解决此问题的，旨在利用数据和多任务标签之间的相关性来减少标签噪声的影响。

该文为解决人脸情感识别的实际案例提供了一些贡献，主要可概括为以下三点：(1)提出了一个带有嘈杂的多任务标签的面部表情识别新问题，该问题的目标是易于获得的廉价多任务注释；(2)提出了一种广义化的公式，在数据和异构多任务标签之间具有明确的联合和边际分布匹配；(3)引入了一种新的对抗学习模型，以基于联合和边际分布的约束条件来优化对情绪预测的训练，这被证明适合于新提出的问题。

带有噪音标签的面部情感识别仅在带有噪音标签的面部图像上训练鲁棒模型。传统的方法是直接用噪声标签分布对噪声建模，但是传统的条件概率建模具有几个明显的缺点，例如转换矩阵缺乏约束条件收敛到真值等。针对于此，本文利用匹配两个联合分布的关键思想，考虑在两对数据和标签上的以下两个联合概率分布：

$$\hat{p}(x, y^0, y^1 = i|\theta) = \hat{p}(x)\hat{p}(y^0, y^1 = i|x, \theta),$$
$$\hat{q}(\tilde{x}, \tilde{y}^0, \tilde{y}^1 = j|\vartheta) = \hat{q}(\tilde{y}^0, \tilde{y}^1 = j)\hat{q}(\tilde{x}|\tilde{y}^0, \tilde{y}^1 = j, \vartheta)$$

由于对现实世界数据的数据分布的显式概率密度函数进行建模难以计算，因此将两个联合分布与精确建模进行匹配通常是不可行的。为克服该问题，本文采用了生成对抗模型方法。其中，编码器的学习函数以从输入图像中推断出干净的标签，解码器的学习函数以生成面部图像，来自嘈杂标签的对应表达式。整体架构如下图所示



为了匹配编码器和解码器捕获的联合分布，在生成器和鉴别器之间进行对抗游戏。鉴别器是专门为匹配面部图像，噪声矢量以及 GY 和 GX 的多任务标签的组的联合分布而设计。对于联合分布对齐，一种自然的方法是将分别从编码器和解码器采样的数据在网络中以进行对抗训练。但是，每个组中的数据是高度异构的，因此直接串联是不合适的。为了减少数据和多任务标签之间的异质性，本文采用多个网络流，并将所有网络流的输出送入网络，完整的目标函数如下，

$$\min_{G_X, G_Y} \max_{D} f(G_Y(x), \tilde{y}) + \lambda(\mathbb{E}_{\hat{p}(x)}[g(D(x, G_Y(x)))] + \mathbb{E}_{\hat{q}(y)}[h(D(G_X(\tilde{y}), \tilde{y}))])$$

文中提出的生成器和鉴别器能够在统一框架内优化基于情绪预测的损失和基于分布匹配的约束。文中根据此方案设计了最小—最大目标函数：

$$\min_{G_X, G_Y} f(G_Y(x), \tilde{y}) + \lambda(\mathbb{E}_{\hat{p}(x)}[\hat{h}(-D(x, G_Y(x)))] + \mathbb{E}_{\hat{q}(y)}[\hat{h}(D(G_X(\tilde{y}), \tilde{y}))])$$

$$\max_{D} \mathbb{E}_{\hat{p}(x)}[g(D(x, G_Y(x)))] + \mathbb{E}_{\hat{q}(y)}[h(D(G_X(\tilde{y}), \tilde{y}))]$$

在该文中，由于将面部情感识别视为目标任务，因此将情感预测用作辅助任务，从而从图像到标签的关系和任务到任务的关系中使目标任务受益，该算法如下图所示。

Algorithm 1 The proposed method

Require: Batch size m, encoder G_Y, decoder G_X, discriminator D, training iterations n, and λ

1: **for** $i \leftarrow 1$ to n **do**
2: Sample data $(x_1, \tilde{y}_1^1, \ldots, \tilde{y}_1^T), \ldots, (x_m, \tilde{y}_m^1, \ldots, \tilde{y}_m^T)$ from the dataset
3: sample Gaussian noise $\tilde{y}_1^0, \ldots, \tilde{y}_m^0$ from $\mathcal{N}(0, 1)$
4: $\tilde{y}_j \leftarrow (\tilde{y}_j^0, \tilde{y}_j^1, \ldots, \tilde{y}_j^T)$ for all j
5: Update G_X, G_Y:
$\min_{G_X, G_Y} \frac{1}{m} \sum_{j=1}^{m} [f(G_Y(x_j), \tilde{y}_j) + \lambda(\hat{h}(-D(x_j, G_Y(x_j))) + \hat{h}(D(G_X(\tilde{y}_j), \tilde{y}_j)))]$
6: Update D:
$\max_D \frac{1}{m} \sum_{j=1}^{m} [g(D(x_j, G_Y(x_j))) + h(D(G_X(\tilde{y}_j, \tilde{y}_j)))]$
7: **end for**

文中在两种情况下对该模型进行评估：(1)用于图像分类的合成噪声标签数据集(CIFAR-10 [25])；(2)用于面部表情识别的两个实用的面部表情数据集(RAF 和 AffectNet)。

下图为实验 1 的结果，可见采用文中提出的模型使得准确率得到提高。

Table 1: Test accuracy on CIFAR-10 synthetic dataset.

Training data	Model	Test acc (%)
Clean labels	VGGNet [46]	88.55
Multi noisy labels	VGGNet-major vote	82.36
	VGGNet	80.23
	AIR [1]	76.37
	SCE [54]	86.34
	Co-teaching [16]	84.21
	LTNet [57]	87.23
	Proposed	**87.90**

下图为基线和在训练步骤中提出的模型的测试准确性曲线的可视化呈现。

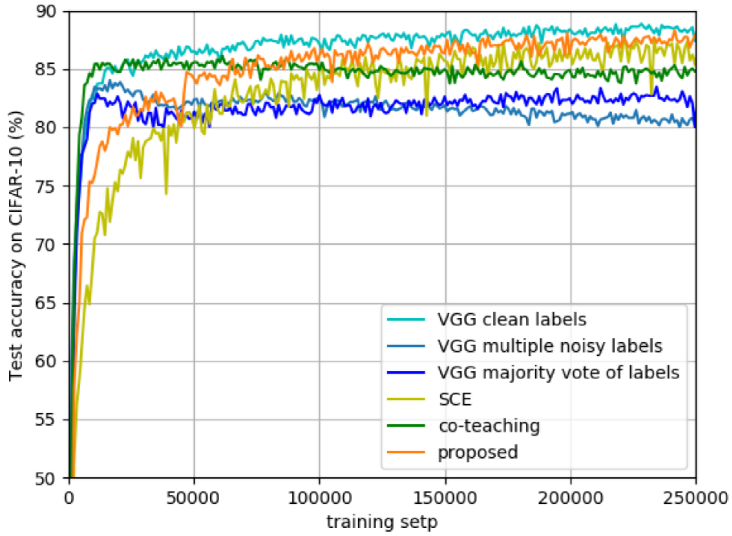


Figure 3: Test accuracy vs. training steps on CIFAR-10 synthetic noisy dataset.

下图为实验 2 的面部情绪数据集的评估结果，可知在多任务情况下，运用本文提出的模型获得的预测准确性更高。

Table 2: Evaluation results on facial emotion datasets. Single-task refers to models trained only with categorical expression labels or valence-arousal labels, and multi-task refers to models trained with both expression and valence-arousal labels. Emotion/Acc (%) denotes test accuracy of categorical expression prediction, the higher the better. VA/CCC, VA/MSE denote CCC and MSE metrics of valence-arousal prediction respectively, the higher the better for CCC and the lower the better for MSE. (**Bold**: best, Underline: second best)

Setting	Model	RAF-base	AffectNet-base		
Task/Metric		Expression/Acc (%)	Expression/Acc (%)	VA/CCC	VA/MSE
Single-task	VGGNet [46]	72.64	43.42	0.6254	0.1438
	SCE [54]	73.96	42.87	-	-
	Co-teaching [16]	75.43	42.36	-	-
	Proposed	74.02	44.52	0.6354	0.1284
Multi-task	VGGNet	73.15	43.36	0.6263	0.1354
	Proposed	**76.10**	**46.08**	**0.6727**	**0.1248**

本文介绍了一个带有噪声的多任务注释的面部情绪识别的问题，在减少人为多任务学习

的标签工作方面具有很大的应用潜力。文中从联合分配匹配的角度介绍了一种新的公式，按照该公式，采用一种新的对抗学习方法来共同优化情绪预测和联合分布学习。最后研究了合成噪声标签数据集和实用的噪声多任务数据库的建立，并通过对它们的评估证明了该方法在解决新问题方面的明显优势。

2) THIN: THrowable Information Networks and Application for Facial Expression Recognition In The Wild

摘要

对于使用深度学习技术解决的许多任务，可以识别一个外生变量，该变量会影响到不同类的外观，并且理想分类器能够对此变量始终保持不变。本文提出了双重外生/内生表示法。文中设计了一个预测层，该预测层使用由外生表示条件限定的深度整体，可以学习自适应的弱预测变量的权重，并且显式地建模外生变量和预测任务之间的依赖关系。此外，文中提出了外源性消除损失的计算，以从内源性表示中删除外源性信息。因此，外生信息被使用了两次，第一次是作为目标任务的条件变量，第二次是在内生表示中产生不变性。本文将该方法命名为 THIN，代表 THrowable Information Net-works。本文在几种可以识别外源信息的情况下，通过实验验证了 THIN，例如大旋转下的数字识别和多尺度下的形状识别。还将其应用于以身份为外生变量的 FER。特别是证明了 THIN 在某些具有挑战性的数据集上的性能明显优于最新方法。

深度学习技术在计算机视觉的监督学习中取得了重大进展，允许共同学习一种表示形式和基于这种表示形式的预测变量。完善的深度学习技术构成了大多数计算机视觉问题中的最新方法，例如对象分类或检测，语义分割或面部和身体分析。然而，在许多此类任务中，对象的外观会受到外生变量的严重影响，理想情况下，任务预测应根据该变量进行不变。

但是，与此同时，从预测系统的角度来看，无论外在变量（例如受试者身份）的变化如何，都应该预测我们的目标任务（例如面部表情）。因此，本文认为与任务相关的表示（称为内生表示）应包含尽可能少的有关外生变量的信息。

综上所述，在这种情况下，该外生变量是数据变化的重要来源，同时也是信息的来源，从该信息中，预测变量的输出应尽可能不变。因此，我们建议使用单独的外在和内在表示。

本文的贡献：(1)提出了一个外生树状深度集成方法，该模型使用内生和外生双重网络。第一个输出表示用于预测任务，而第二个输出的表示通过适应性和联合学习更多相关的弱预测变量，以进行深度相关的调整；(2)提出了一种外源消除损失，通过内源表示与外源表示之间的正交性，从内源表示中消除外源变异；(3)在具有不同外生变量的多个任务上实验性地验证了这种方法。

文中通过深度神经网络对外生信息建模，然后从定义一个简单的基线模型开始，然后逐步引入其他的架构，从而描述如何明确地合并外生表示和任务预测之间的依赖关系，整体架构如下图所示。

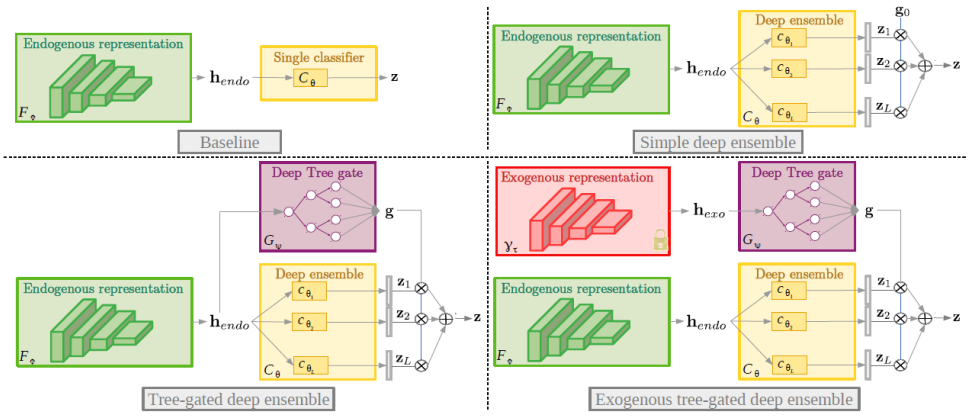


Figure 3: Architectures overview. **Baseline** (top-left): a simple baseline, composed of a stack of representation and classification layers. **Single deep ensemble** (top-right): the classification layer is represented an average of several smaller classifiers. **Tree-gated deep ensemble** (bottom-left): instead of merely averaging the classifiers we learn the mixture weights using a deep differentiable tree. **Exogenous tree-gated deep ensemble** (bottom-right): we use a second representation network related to the exogenous representation, which appears as a better conditioning variable.

如上图所示，主要呈现了基线框架，简单的深度集成方法框架，树状深度集成方法框架，外生树状深度集成方法框架。从基线框架开始，通过自适应加权深层集成的预测并利用外源表示来逐步改进框架的设计方法。

树状深度集成网络通过参数$\emptyset$，θ，φ优化相应的损失，然后将与外生变量有关的信息分解为内生表示中的任务，并将提取的外生和内生的特征输入网络$F_{\emptyset}$和γ_{τ}进行输出，通过超参数λ进行实验设置，从而实现从内在表征中去除外源性信息。

$$\mathscr{L}(\theta,\phi,\psi)=\mathscr{L}_{sup}(\theta,\phi,\psi)+\lambda\mathscr{L}_{sim}(\phi)$$

文中通过将模型在合成数据集上进行评估，从中可以清楚地识别外生变量。紧接着，在真实的 FER 数据集中定性和定量验证模型，主要是介绍了用于训练或测试所提出方法的数据集，具体的实现细节。下图中 Table 2 为在 MNIST-R 和 dSprites 数据集上，根据平均准确度比较不同体系结构；Figure 4 为 MNIST-R 以平均准确度表示的消融外源表征消除的消融研究

Table 2: Comparison of different architectures in term of average accuracy (%) on MNIST-R and dSprites databases. †: oracle classifier (tree-gated deep ensemble conditionned by the ground truth rotation).

Method	MNIST-R	dSprites
Baseline	96.83	96.53
Simple deep ensemble	96.81	92.68
Tree-gated deep ensemble	97.31	95.91
Exogenous tree-gated deep ensemble	98.07	98.1
Oracle†	98.06	98.43
THIN	**98.26**	**98.5**

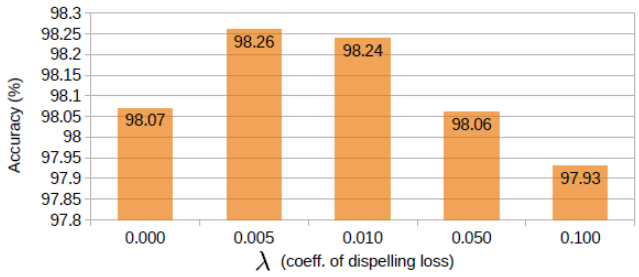


Figure 4: Ablation study for the exogenous representation-dispelling loss in term of average accuracy (%) on MNIST-R.

除了在 MNIST 数据集上之外，文中还在 RAF-DB，AffectNet 和 ExpW 数据集上进行了实验验证，Table 3 从平均准确率上比较了不同的体系架构，Figure 5 是在数据集 RAF-DB 上进行消融研究的结果。

Table 3: Comparison of different architectures in term of average accuracy (%) on RAF-DB, AffectNet and ExpW databases.

Method	RAF-DB	AffectNet	ExpW
Baseline (VGG16)	82.99	61.31	70.96
Baseline (VGGFace)	84.06	61.66	71.57
Simple deep ensemble	85.59	63.00	75.05
Tree-gated deep ensemble	86.38	63.34	75.17
Exogenous tree-gated deep ensemble	87.29	63.71	75.74
THIN	**87.81**	**63.97**	**76.08**

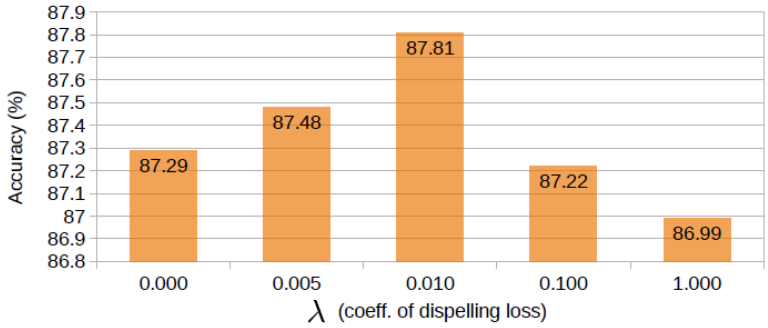


Figure 5: Ablation study for identity-expression disentangling in term of average accuracy (%) on RAF-DB.

最后将 THIN 与最新的 FER 方法进行了比较，证明了 THIN 在当今最新的，具有挑战

性的 FER 数据库上的性能明显优于最新技术。

Table 4: Comparison with state-of-the-art approaches in term of accuracy (%).

Method	RAF-DB	AffectNet	ExpW
PG-CNN [35]	83.27	55.33	-
Separate loss [34]	86.38	58.89	-
IPA2LT [58]	86.77	57.31	-
RAN [54]	86.9	59.5	-
Covariance pooling [1]	87.00	-	-
SNA [17]	-	62.7	-
BReG-Net [18]	-	63.54	-
PAT-VGG [7]	86.28	-	71.5
EAFR [36]	82.69	-	71.90
THIN	**87.81**	**63.97**	**76.08**

本文中所提出的模型具有较多的应用可能性。首先，理论上可以将 THIN 直接应用于其他问题，例如以姿势或比例作为外生变量的身体姿势估计，或具有领域信息的语义分割。其次，在本文中仅使用一个外生变量来训练 THIN。但是，可以尝试使用多个这样的变量和表示网络以及某种融合方案来应用。此外可以尝试使用身份作为外生变量的 THIN 来预测面部表情，然后使用以面部表情作为外生变量的另一个 THIN 来预测身份，依此类推，以迭代地完善 FER 和身份预测。

发布链接：https://mp.weixin.qq.com/s/gF6tB68qPJjf5cwtkRkYiA

Reference

1. 基于深度学习的自然场景下多人脸检测
2. Facial Emotion Recognition with Noisy Multi-task Annotations
3. THIN: THrowable Information Networks and Application for Facial Expression Recognition in the Wild