Transferability and Hardness of Supervised Classification Tasks
— Supplemental material —

Anh T. Tran* VinAI Research anstar1111@gmail.com

Cuong V. Nguyen Amazon Web Services nguycuo@amazon.com

Tal Hassner* Facebook AI talhassner@gmail.com

This supplemental material includes a full proof of Theorem 1, more details on task hardness, technical implementation details, additional results for multi-class classification, and full transferability results on CelebA and AwA2 (omitting transferability plots for the 312 tasks in the CUB dataset, due to space requirements). We note that some concessions were made in order to fit these results to the limited format of an ICCV supplemental document.

1. Proof of theorem 1

From the definition of $\widetilde{\mathrm{Trf}}(T^Z \to T^Y)$, we have:

$$\begin{aligned}
\widetilde{\mathrm{Trf}}(T^Z \to T^Y) &= l_Y(w_Z, k_Y) && \text{(definition of } \widetilde{\mathrm{Trf}}\text{)} \\
&\geq l_Y(w_Z, \bar{k}) && \text{(definition of } k_Y \text{ and } \bar{k} \in K\text{)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{z \in \mathcal{Z}} \hat{P}(y_i|z) P(z|x_i; w_Z, h_Z) \right) && \text{(construction of } \bar{k}\text{)} \\
&\geq \frac{1}{n} \sum_{i=1}^{n} \log \left(\hat{P}(y_i|z_i) P(z_i|x_i; w_Z, h_Z) \right) && \text{(replacing the sum by one of its elements)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \log \hat{P}(y_i|z_i) + \frac{1}{n} \sum_{i=1}^{n} \log P(z_i|x_i; w_Z, h_Z). && (1)
\end{aligned}$$

Note that the second term in Eq. (1) is:

$$\frac{1}{n} \sum_{i=1}^{n} \log P(z_i|x_i; w_Z, h_Z) = l_Z(w_Z, h_Z). \quad (2)$$

Furthermore, the first term in Eq. (1) is:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \log \hat{P}(y_i|z_i) &= \frac{1}{n} \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} \left(\sum_{i \,:\, y_i = y \text{ and } z_i = z} \log \hat{P}(y|z) \right) && \text{(group the summands by values of } y_i \text{ and } z_i\text{)} \\
&= \frac{1}{n} \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} \left(|\{i : y_i = y \text{ and } z_i = z\}| \;\log \hat{P}(y|z) \right) && \text{(by counting)} \\
&= \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} \left(\frac{|\{i : y_i = y \text{ and } z_i = z\}|}{n} \;\log \hat{P}(y|z) \right) \\
&= \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} \left(\hat{P}(y, z) \;\log \frac{\hat{P}(y, z)}{\hat{P}(z)} \right) && \text{(definitions of } \hat{P}(y, z) \text{ and } \hat{P}(y|z)\text{)} \\
&= -H(Y|Z). && (3)
\end{aligned}$$

*Work at Amazon Web Services, prior to joining current affiliation.

From Eq. (1), (2), and (3), we have $\widetilde{\mathrm{Trf}}(T^Z \to T^Y) \geq l_Z(w_Z, h_Z) - H(Y|Z)$. Hence, the theorem holds.

2. More details on task hardness

On the definition of task hardness. In our paper, we assume non-overfitting of trained models. When train and test sets are sampled from the *same distribution*, this assumption typically holds for appropriately trained models [14]. This property also shows that our definition of hardness, Eq. (13) in the main paper, does not conflict with the results of Zhang et al. [14]: In such cases, the training loss of Eq. (13) correlates with the test error, and thus this definition indeed reflects task hardness, explaining the relationships between train and test errors observed in our hardness results.

On the representation for trivial tasks. *Any* representation for a trivial source task can fit the constant label perfectly (zero training loss). In theory, if we choose the optimal w_Z (in Eq. (1) of the main paper) as our representation, we can show Eq. (14) in the main paper. In practice, of course we cannot infer the optimal w_Z from the trivial source task, but Eq. (14) shows that we can still connect it to $H(Z|C)$.

3. Technical implementation details

Computing the CE. Computing the CE is straightforward and involves the following steps:

1. Loop through the training labels of both tasks T^Z and T^Y and compute the empirical joint distribution $\hat{P}(y, z)$ by counting (Eq. (6) in the paper).
2. Loop through the training labels again and compute the CE using Eq. (3) above. That is,

$$H(Y|Z) = -\frac{1}{n}\sum_{i=1}^{n} \log \hat{P}(y_i|z_i).$$

Thus, computing the CE only requires running two loops through the training labels. This process is computationally efficient. In the most extreme case, computing the transferability of face recognition ($|\mathcal{Z}| > 10k$) to a facial attribute, with $|\mathcal{Y}| = 2$, required *less than a second* on a standard CPU.

This run time should be compared with the *hours (or days)* required to train deep models in order to empirically measure transferability following the process described by previous work. In particular, Taskonomy [13] reported *over 47 thousand hours of GPU runtime* in order to establish relationships between their 26 tasks.

Dedicated attribute training. Given a source task T^Z, we train a dedicated CNN for this task with standard ResNet-18 V2 implemented in the MXNet deep learning library [1].[1] We set the initial learning rate to 0.01. Learning rate was then divided by 10 after each 12 epochs. Training converged in less than 40 epochs in all 437 tasks.

Task transfer with linear SVM. After training a deep representation for a source task T^Z, we transfer it to a target task T^Y using linear support vector machines (lSVM).

First, we use the trained CNN, denoted in the paper as w_Z, to extract deep embeddings for the entire training data (one embedding per input image. Each embedding is a vector $r \in \mathbb{R}^{2048}$, which we obtain from the penultimate layer of the network. We then use these embeddings, along with the corresponding labels for target task, T^Y, to train a standard lSVM classifier, implemented by SK-Learn [8]. The lSVM parameters were kept unchanged from their default values.

Given unseen testing data, we first extract their embeddings with w_Z. We then apply the trained lSVM classifier on these features to predict labels for target task, T^Y.

4. Additional results: Generalization to multi-class

Transferability generalizes well to multi-class, as evident in our face recognition (10k labels)-to-attribute tests in Sec. 5.2. Table 1 below reports hardness tests with multi-class, CelebA, attribute aggregates. Generally speaking, the harder the task, the lower the accuracy obtained.

[1] Model available from: https://mxnet.apache.org/api/python/gluon/model_zoo.html.

Multi-class	Straight/Wavy/Other	Black/Blonde/Other	Arched/Bushy/Other
Hardness ↓	1.040	0.925	0.867
Dedicated Res18	0.713	0.859	0.797
Multi-class	Bangs/Receding/Other	Gray/Blonde/Other	Goatee/Beard/None
Hardness ↓	0.690	0.575	0.557
Dedicated Res18	0.900	0.943	0.937

Table 1. Multi-class hardness examples on CelebA data.

5. Full transferability results

5.1. Attribute prediction on CelebA [7]

Figure 1. **Attribute prediction; CE vs. test errors on CelebA (Extended from Fig. 2(a-d) in the paper).** The source attribute, T^Z, in each plot is named in the plot title. Points represent different target tasks T^Y. Corr is the Pearson correlation coefficient between the two variables and p is the statistical significance of the correlation. In all cases, the correlation is statistically significant.

5.2. CelebA: Transferability from identity to attributes

	Attribute:	Male	Bald	Gray Hair	Mustache	Double Chin	Chubby	Sideburns	Goatee	Young	Wear Hat	
1	LNets+ANet 2015 [7]	0.980	0.980	0.970	0.950	0.920	0.910	0.960	0.950	0.870	0.990	
2	Walk and Learn 2016 [10]	0.960	0.920	0.950	0.900	0.930	0.890	0.920	0.920	0.860	0.960	
3	MOON 2016 [9]	0.981	0.988	0.981	0.968	0.963	0.954	0.976	0.970	0.881	0.990	
4	LMLE 2016 [5]	0.990	0.900	0.910	0.730	0.740	0.790	0.880	0.950	0.870	0.990	
5	CR-I 2017 [2]	0.960	0.970	0.950	0.940	0.890	0.870	0.920	0.960	0.840	0.980	
6	MCNN-AUX 2017 [4]	0.982	0.989	0.982	0.969	0.963	0.957	0.978	0.972	0.885	0.990	→
7	DMTL 2018 [3]	0.980	0.990	0.960	0.970	0.990	0.970	0.980	0.980	0.900	0.990	
8	Face-SSD 2019 [6]	0.973	0.986	0.976	0.960	0.960	0.951	0.966	0.963	0.876	0.985	
9	Conditional Entropy↑	0.017	0.026	0.052	0.062	0.083	0.087	0.088	0.089	0.095	0.107	
10	Dedicated Res18	0.985	0.990	0.980	0.968	0.959	0.951	0.976	0.974	0.879	0.991	
11	FromID SVM	0.992	0.991	0.981	0.968	0.963	0.957	0.976	0.973	0.899	0.988	

	Attribute:	Eye glasses	Pale Skin	Wear Necktie	Blurry	No Beard	Receding Hairline	5 clock Shadow	Rosy Cheeks	Blond Hair	Big Lips	
1		0.990	0.910	0.930	0.840	0.950	0.890	0.910	0.900	0.950	0.680	
2		0.970	0.850	0.840	0.910	0.900	0.840	0.840	0.960	0.920	0.780	
3		0.995	0.970	0.966	0.957	0.956	0.936	0.940	0.948	0.959	0.715	
4		0.980	0.800	0.900	0.590	0.960	0.760	0.820	0.780	0.990	0.600	
5		0.960	0.920	0.880	0.850	0.940	0.870	0.900	0.880	0.950	0.680	
6		0.996	0.970	0.965	0.962	0.960	0.938	0.945	0.952	0.960	0.715	→
7		0.990	0.970	0.970	0.960	0.970	0.940	0.950	0.960	0.910	0.880	
8		0.992	0.957	0.956	0.950	0.949	0.931	0.929	0.943	0.936	0.778	
9		0.109	0.122	0.131	0.139	0.141	0.141	0.145	0.152	0.16	0.161	
10		0.997	0.970	0.963	0.963	0.961	0.936	0.942	0.950	0.961	0.715	
11		0.996	0.958	0.941	0.956	0.958	0.933	0.937	0.939	0.949	0.710	

	Attribute:	Bushy Eyebrows	Wear Lipstick	Big Nose	Bangs	Narrow Eyes	Wear Necklace	Heavy Makeup	Black Hair	Wear Earrings	Arched Eyebrows	
1		0.900	0.930	0.780	0.950	0.810	0.710	0.900	0.880	0.820	0.790	
2		0.930	0.920	0.910	0.960	0.790	0.770	0.960	0.840	0.910	0.870	
3		0.926	0.939	0.840	0.958	0.865	0.870	0.910	0.894	0.896	0.823	
4		0.820	0.990	0.800	0.980	0.590	0.590	0.980	0.920	0.830	0.790	
5		0.840	0.940	0.800	0.950	0.720	0.740	0.840	0.900	0.830	0.800	
6		0.928	0.941	0.845	0.960	0.872	0.866	0.915	0.898	0.904	0.834	→
7		0.850	0.930	0.920	0.960	0.900	0.890	0.920	0.850	0.910	0.860	
8		0.896	0.926	0.823	0.952	0.890	0.878	0.907	0.879	0.869	0.820	
9		0.192	0.202	0.232	0.236	0.252	0.252	0.27	0.286	0.291	0.306	
10		0.927	0.935	0.828	0.961	0.875	0.859	0.916	0.901	0.896	0.834	
11		0.919	0.940	0.845	0.950	0.863	0.865	0.897	0.869	0.853	0.822	

	Attribute:	Brown Hair	Bags U Eyes	Oval Face	Straight Hair	Pointy Nose	Attractive	Wavy Hair	High Cheeks	Smiling	Mouth Open	Average (all)
1		0.800	0.790	0.660	0.730	0.720	0.810	0.800	0.870	0.920	0.920	0.873
2		0.810	0.870	0.790	0.750	0.770	0.840	0.850	0.950	0.980	0.970	0.887
3		0.894	0.849	0.757	0.823	0.765	0.817	0.825	0.870	0.926	0.935	0.909
4		0.870	0.730	0.680	0.730	0.720	0.880	0.830	0.920	0.990	0.960	0.838
5		0.860	0.800	0.660	0.730	0.730	0.830	0.790	0.890	0.930	0.950	0.866
6		0.892	0.849	0.758	0.836	0.775	0.831	0.839	0.876	0.927	0.937	0.913
7		0.960	0.990	0.780	0.850	0.780	0.850	0.870	0.880	0.940	0.940	0.926
8		0.835	0.825	0.748	0.834	0.749	0.813	0.851	0.868	0.918	0.919	0.903
9		0.315	0.324	0.339	0.339	0.341	0.361	0.381	0.476	0.521	0.551	
10		0.886	0.834	0.752	0.836	0.769	0.823	0.842	0.878	0.933	0.943	0.911
11		0.854	0.838	0.733	0.812	0.769	0.820	0.800	0.859	0.909	0.901	0.902

Table 2. **Transferability from face recognition to facial attributes. (Extended from Table 1 in the paper)** Results for CelebA attributes, sorted in ascending order of row 9 (decreasing transferability). Classification accuracies are shown for all 40 attributes. Subject specific attributes, e.g., *male* and *bald*, are more transferable than expression related attributes such as *smiling* and *mouth open*. These identity specific attributes corresponds to the automatic grouping presented in the original CelebA paper [7]. Unlike them, however, we obtain this grouping without necessitating the training of a deep attribute classification model. Unsurprisingly, transfer results (row 11) are best on these subject specific attributes and worst for less related attributes. Rows 1-8 provide published state of the art results. Despite training only an lSVM for attribute, row 11 results are comparable with more elaborate attribute classification systems. For details, see Sec. 5.2.

5.3. Attribute prediction on AwA2 [12]

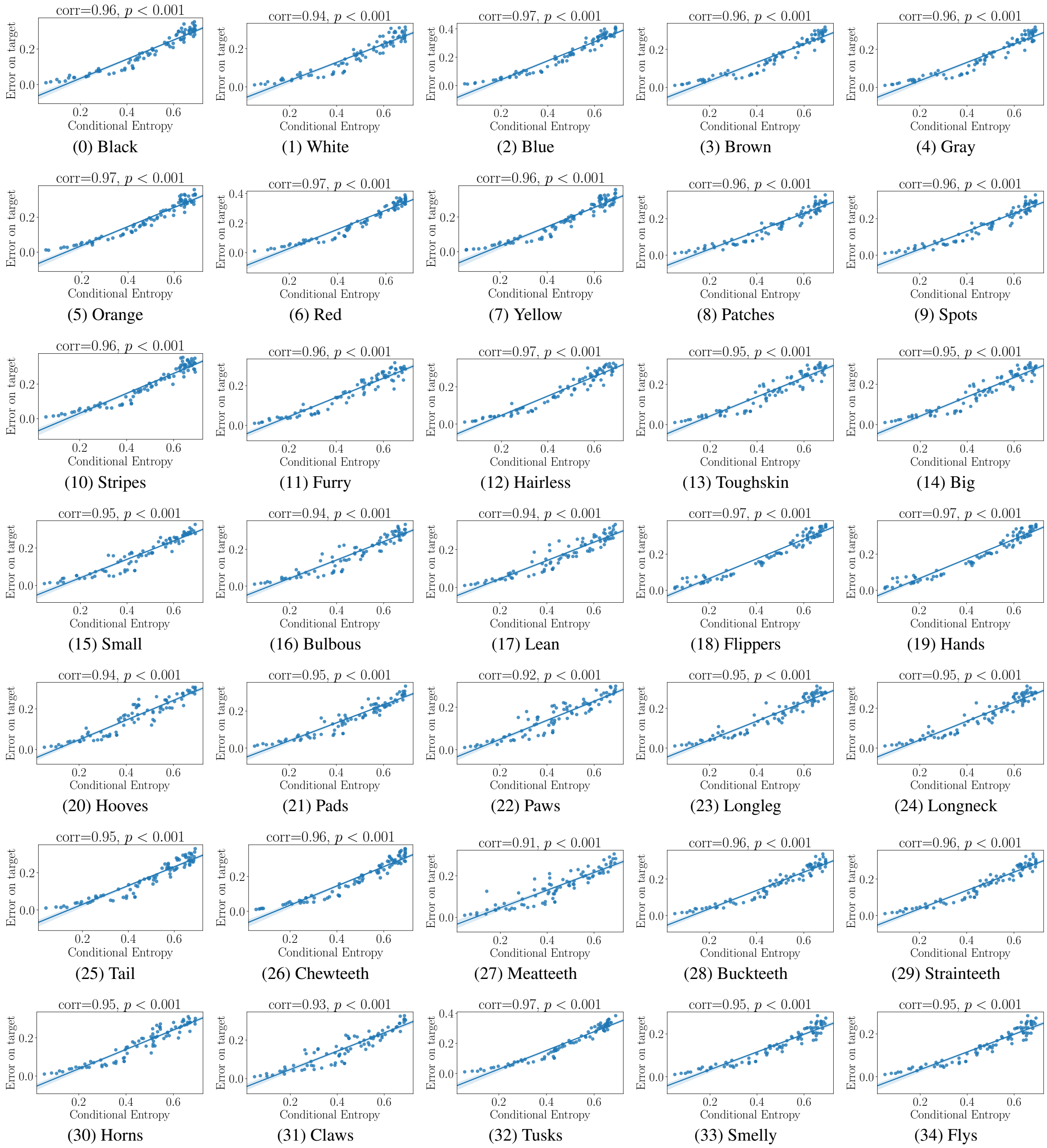
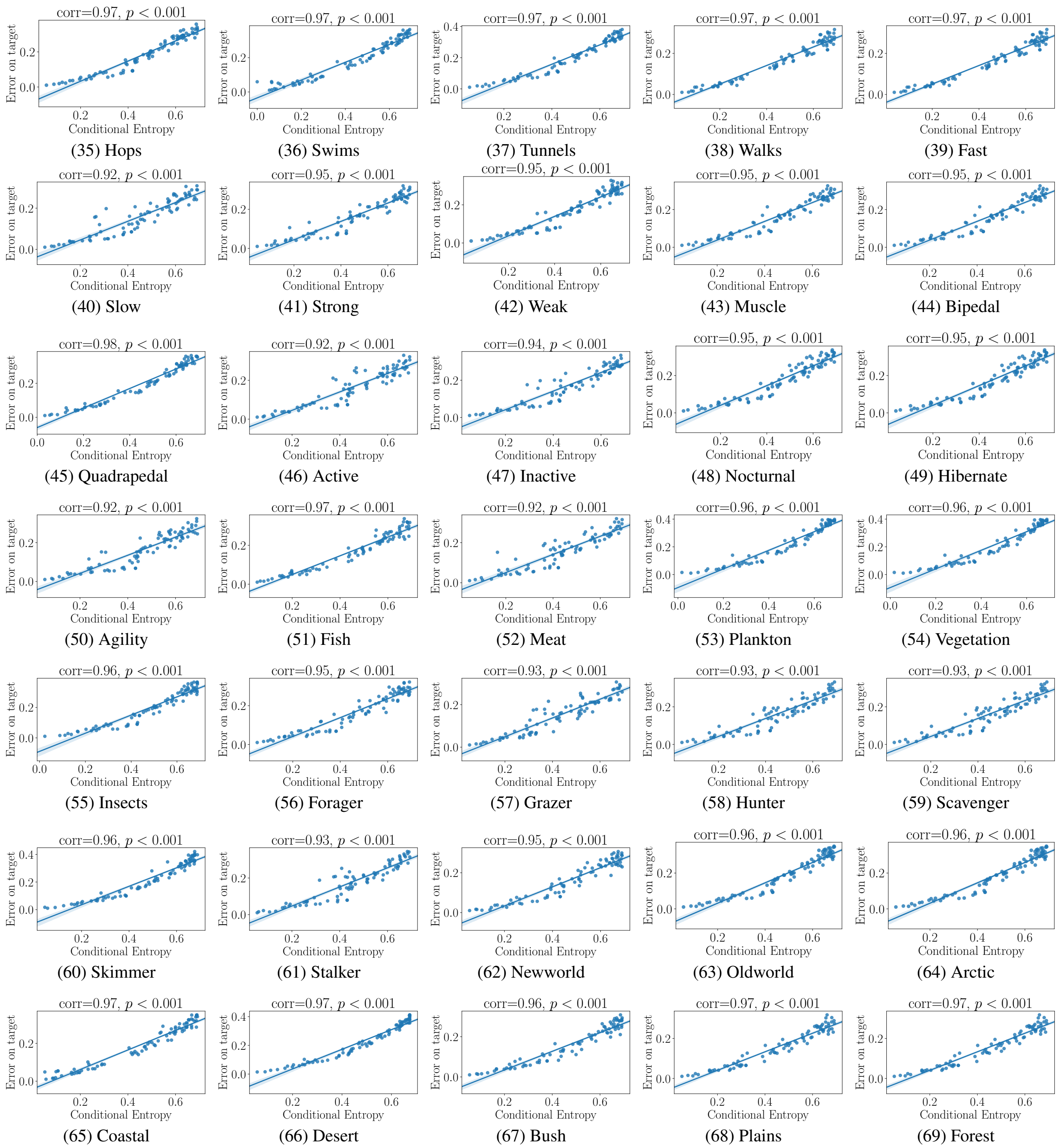


Figure 2. **Attribute prediction; CE vs. test errors on AwA2 (Extended from Fig. 2(e-h) in the paper; part 1).** The source attribute, T^Z, in each plot is named in the plot title. Points represent different target tasks T^Y. Corr is the Pearson correlation coefficient between the two variables and p is the statistical significance of the correlation. In all cases, the correlation is statistically significant.



(35) Hops (36) Swims (37) Tunnels (38) Walks (39) Fast

(40) Slow (41) Strong (42) Weak (43) Muscle (44) Bipedal

(45) Quadrapedal (46) Active (47) Inactive (48) Nocturnal (49) Hibernate

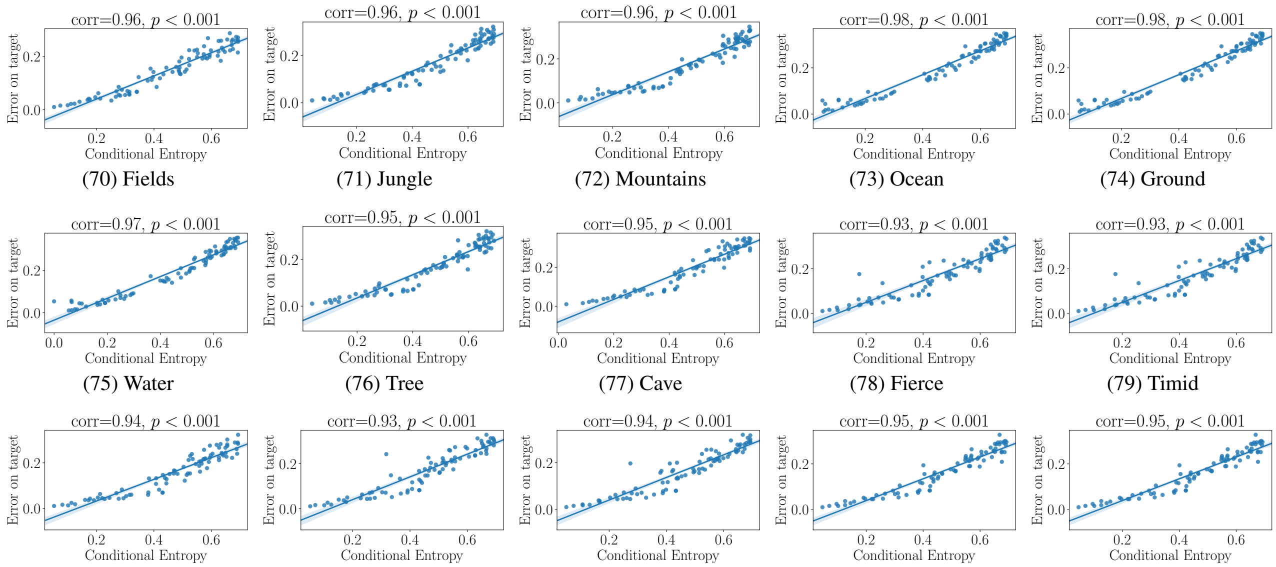
(50) Agility (51) Fish (52) Meat (53) Plankton (54) Vegetation

(55) Insects (56) Forager (57) Grazer (58) Hunter (59) Scavenger

(60) Skimmer (61) Stalker (62) Newworld (63) Oldworld (64) Arctic

(65) Coastal (66) Desert (67) Bush (68) Plains (69) Forest

Figure 3. **Attribute prediction; CE vs. test errors on AwA2 (Extended from Fig. 2(e-h) in the paper; part 2).** The source attribute, T^Z, in each plot is named in the plot title. Points represent different target tasks T^Y. Corr is the Pearson correlation coefficient between the two variables and p is the statistical significance of the correlation. In all cases, the correlation is statistically significant.



(70) Fields (71) Jungle (72) Mountains (73) Ocean (74) Ground

(75) Water (76) Tree (77) Cave (78) Fierce (79) Timid

(80) Smart (81) Group (82) Solitary (83) Nestspot (84) Domestic

Figure 4. **Attribute prediction; CE vs. test errors on AwA2 (Extended from Fig. 2(e-h) in the paper; part 3).** The source attribute, T^Z, in each plot is named in the plot title. Points represent different target tasks T^Y. Corr is the Pearson correlation coefficient between the two variables and p is the statistical significance of the correlation. In all cases, the correlation is statistically significant.

6. Full hardness results

6.1. CelebA [7] attribute prediction hardness

1 Attribute	Bald	Mustache	Gray Hair	Pale Skin	Double Chin	Wearing Hat	Blurry	Sideburns	Chubby	Goatee	→
2 Conditional Entropy↑	0.107	0.173	0.174	0.177	0.189	0.194	0.201	0.217	0.22	0.235	→
→	Eyeglasses	Rosy Cheeks	Wearing Necktie	Receding Hairline	5 oClock Shadow	Narrow Eyes	Wearing Necklace	Bushy Eyebrows	Blond Hair	Bangs	→
→	0.241	0.242	0.261	0.278	0.349	0.357	0.373	0.409	0.419	0.425	→
→	No Beard	Wearing Earrings	Bags Under Eyes	Brown Hair	Straight Hair	Young	Big Nose	Black Hair	Big Lips	Arched Eyebrows	→
→	0.448	0.485	0.507	0.508	0.512	0.535	0.545	0.55	0.552	0.58	→
→	Pointy Nose	Oval Face	Wavy Hair	Heavy Makeup	Male	High Cheekbones	Wearing Lipstick	Smiling	Mouth Slightly Open	Attractive	
→	0.591	0.597	0.627	0.667	0.679	0.689	0.692	0.693	0.693	0.693	

Table 3. **CelebA task hardness.** CelebA facial attributes sorted in ascending order of hardness along with their respective hardness scores. Hardness scores listed above are compared with empirical test errors for each task and shown to be strongly correlated (Fig. 6(a) in the paper). Note that the *male* classification task, appearing here as relatively hard, is the easiest task to transfer from face recognition (Table 2).

6.2. AwA2 [12] attribute prediction hardness

1 Attribute	Flys	Red	Skimmer	Desert	Plankton	Insects	Tunnels	Hands	Tusks	Strainteeth	Cave	Blue	Stripes	Scavenger	Hops	→
2 Conditional Entropy↑	0.057	0.089	0.112	0.125	0.140	0.170	0.181	0.200	0.205	0.229	0.243	0.251	0.263	0.270	0.284	→
→	Oldworld	Orange	Yellow	Quadrapedal	Flippers	Ground	Ocean	Coastal	Arctic	Walks	Swims	Water	Weak	Longneck	Bipedal	→
→	0.294	0.303	0.315	0.324	0.345	0.381	0.391	0.392	0.408	0.429	0.433	0.433	0.439	0.444	0.444	→
→	Tree	Chewteeth	Hibernate	Nocturnal	Fast	Furry	Stalker	Newworld	Tail	Horns	Hairless	Jungle	Buckteeth	Spots	Active	→
→	0.450	0.454	0.477	0.487	0.501	0.507	0.511	0.514	0.515	0.518	0.524	0.526	0.527	0.562	0.570	→
→	Mountains	Strong	Bush	Pads	Fish	Big	Timid	Hunter	Small	Brown	Longleg	Hooves	Agility	Nestspot	Smart	→
→	0.580	0.582	0.585	0.593	0.599	0.601	0.617	0.624	0.630	0.643	0.644	0.645	0.646	0.648	0.648	→
→	Group	Meat	Patches	Fierce	Forest	Claws	Black	Muscle	Meatteeth	Slow	Fields	Vegetation	Domestic	Grazer	Gray	→
→	0.649	0.651	0.656	0.661	0.667	0.667	0.669	0.672	0.674	0.675	0.678	0.679	0.682	0.686	0.690	→
→	Paws	Plains	Solitary	Toughskin	Bulbous	White	Forager	Inactive	Smelly	Lean						
→	0.691	0.692	0.692	0.692	0.693	0.693	0.693	0.693	0.693	0.693						

Table 4. **AWA2 task hardness.** AWA2 attributes sorted in ascending order of hardness along with their respective hardness scores. Hardness scores listed above are compared with empirical test errors for each task and shown to be strongly correlated (Fig. 6(b) in the paper).

6.3. CUB [11] attribute prediction hardness

1	Attribute	ec::purple	lc::green	ec::green	ec::olive	blc::green	blc::purple	brc::purple	ec::pink	bec::purple	sh::owl-like	→
2	Conditional Entropy↑	0.011	0.015	0.015	0.016	0.017	0.019	0.022	0.022	0.025	0.025	→
	→	lc::olive	unc::purple	fc::pink	upc::pink	untc::purple	cc::pink	wc::pink	bec::pink	tc::purple	lc::iridescent	→
	→	0.025	0.027	0.027	0.028	0.030	0.030	0.031	0.032	0.033	0.034	→
	→	tc::pink	wc::purple	brc::pink	uptc::purple	blc::olive	untc::pink	bkc::pink	lc::purple	cc::purple	pc::pink	→
	→	0.035	0.035	0.035	0.035	0.036	0.036	0.036	0.037	0.037	0.038	→
	→	nc::pink	unc::pink	nc::purple	bkc::purple	blc::pink	fc::purple	upc::purple	ec::blue	pc::purple	uptc::pink	→
	→	0.038	0.040	0.040	0.041	0.042	0.042	0.042	0.049	0.051	0.053	→
	→	tc::green	fc::green	bkc::rufous	uptc::rufous	blc::iridescent	sh::long-legged-like	bec::rufous	untc::rufous	ec::rufous	brc::green	→
	→	0.054	0.054	0.057	0.057	0.057	0.061	0.064	0.064	0.064	0.064	→
	→	wc::rufous	lc::rufous	cc::green	upc::rufous	untc::green	bec::green	tc::olive	brc::rufous	blc::rufous	unc::green	→
	→	0.066	0.067	0.069	0.070	0.071	0.072	0.072	0.074	0.074	0.074	→
	→	tc::rufous	brc::iridescent	nc::green	pc::rufous	unc::rufous	uptc::green	nc::rufous	bec::iridescent	ec::buff	tc::iridescent	→
	→	0.074	0.075	0.076	0.076	0.076	0.077	0.077	0.080	0.081	0.082	→
	→	bls::hooked	fc::rufous	lc::blue	ec::orange	untc::iridescent	unc::iridescent	uptc::iridescent	untc::orange	bkc::orange	cc::rufous	→
	→	0.082	0.082	0.082	0.084	0.084	0.089	0.089	0.089	0.089	0.090	→
	→	fc::iridescent	ec::yellow	sh::chicken-like-marsh	nc::orange	cc::iridescent	si::very-large	tc::orange	untc::red	pc::iridescent	uptc::orange	→
	→	0.092	0.092	0.093	0.093	0.094	0.094	0.095	0.095	0.095	0.097	→
	→	bkc::green	cc::orange	ec::grey	pc::green	wc::green	lc::yellow	wc::iridescent	brc::olive	upc::green	nc::iridescent	→
	→	0.099	0.099	0.099	0.099	0.100	0.100	0.101	0.103	0.107	0.107	→
	→	bls::specialized	bkc::iridescent	fc::olive	cc::olive	blc::blue	fc::orange	bec::blue	bls::curved	bls::needle	wc::orange	→
	→	0.108	0.108	0.110	0.112	0.112	0.113	0.113	0.113	0.113	0.114	→
	→	bec::olive	lc::pink	wc::red	sh::upwl	unc::olive	uptc::red	bkc::red	ec::red	upc::orange	upc::iridescent	→
	→	0.115	0.117	0.117	0.117	0.118	0.118	0.119	0.120	0.123	0.126	→
	→	upc::red	unc::blue	sh::hawk-like	brc::orange	nc::olive	brc::blue	bec::orange	sh::upland-ground-like	hp::spotted	tc::blue	→
	→	0.129	0.130	0.133	0.134	0.134	0.137	0.139	0.139	0.141	0.141	→
	→	unc::orange	untc::olive	ec::brown	pc::orange	si::large	lc::red	uptc::olive	hp::unique-pattern	bec::red	ec::white	→
	→	0.144	0.145	0.147	0.153	0.155	0.156	0.156	0.156	0.161	0.162	→
	→	pc::red	nc::red	hp::crested	lc::white	blc::red	sh::swallow-like	tc::red	brc::red	bls::spatulate	unc::red	→
	→	0.162	0.162	0.165	0.167	0.167	0.170	0.172	0.173	0.174	0.175	→
	→	fc::red	bkc::olive	pc::olive	wc::olive	hp::masked	untc::blue	sh::hummingbird-like	ts::forked-tail	sh::sandpiper-like	cc::red	→
	→	0.175	0.180	0.181	0.182	0.183	0.184	0.184	0.189	0.190	0.193	→
	→	upc::olive	fc::blue	wc::blue	uptc::blue	nc::blue	blc::white	blc::yellow	bkc::blue	cc::blue	upc::blue	→
	→	0.193	0.205	0.208	0.209	0.212	0.217	0.219	0.220	0.222	0.226	→
	→	tp::spotted	sh::pigeon-like	bll::longer-than-head	uptc::yellow	sh::duck-like	pc::blue	bep::spotted	bls::hooked-seabird	brp::spotted	sh::tree-clinging-like	→
	→	0.229	0.231	0.231	0.231	0.231	0.235	0.240	0.240	0.248	0.250	→
	→	sh::gull-like	hp::striped	blc::orange	cc::yellow	untc::yellow	bkp::spotted	blc::brown	wp::spotted	nc::yellow	ws::long-wings	→
	→	0.254	0.260	0.262	0.271	0.278	0.284	0.293	0.294	0.294	0.294	→
	→	tc::brown	bkc::yellow	fc::yellow	ws::broad-wings	lc::brown	hp::malar	wc::yellow	bec::brown	lc::orange	ts::fan-shaped-tail	→
	→	0.298	0.301	0.307	0.313	0.318	0.318	0.318	0.323	0.328	0.330	→
	→	ts::squared-tail	upc::yellow	unc::brown	bep::striped	tc::yellow	ec::black	hp::capped	hp::eyeline	hp::eyebrow	cc::white	→
	→	0.336	0.346	0.356	0.360	0.363	0.368	0.371	0.374	0.375	0.380	→
	→	brc::brown	ws::tapered-wings	brp::striped	fc::buff	cc::buff	bls::dagger	ts::rounded-tail	fc::white	pc::yellow	brc::yellow	→
	→	0.381	0.382	0.383	0.384	0.386	0.402	0.403	0.412	0.416	0.421	→
	→	tc::buff	blc::buff	uptc::buff	bec::yellow	untc::buff	bec::black	hp::eyering	bep::multi-colored	unc::yellow	tc::grey	→
	→	0.425	0.431	0.437	0.440	0.443	0.444	0.446	0.450	0.450	0.451	→
	→	nc::buff	tp::striped	uptc::white	fc::brown	bkc::buff	lc::buff	nc::brown	bec::grey	upc::buff	pc::buff	→
	→	0.456	0.466	0.476	0.483	0.483	0.483	0.484	0.485	0.490	0.490	→
	→	brc::buff	si::medium	bkc::white	bec::buff	unc::black	brp::multi-colored	wc::buff	cc::brown	brc::grey	unc::buff	→
	→	0.492	0.493	0.494	0.494	0.496	0.496	0.498	0.501	0.503	0.510	→
	→	unc::grey	bkp::striped	fc::grey	untc::brown	brc::black	uptc::brown	ts::pointed-tail	cc::grey	tc::black	si::very-small	→
	→	0.513	0.514	0.518	0.520	0.528	0.530	0.531	0.534	0.544	0.549	→
	→	nc::white	untc::white	bkp::multi-colored	pc::brown	nc::grey	bkc::brown	upc::white	untc::grey	wp::striped	hp::plain	→
	→	0.549	0.554	0.555	0.564	0.572	0.573	0.582	0.584	0.587	0.588	→
	→	bls::cone	uptc::grey	blc::grey	wc::white	upc::brown	pc::white	nc::black	pc::grey	wc::brown	tp::multi-colored	→
	→	0.590	0.591	0.594	0.596	0.599	0.602	0.604	0.605	0.606	0.606	→
	→	lc::black	bkc::grey	ws::pointed-wings	lc::grey	wp::solid	wp::multi-colored	wc::grey	upc::grey	fc::black	bep::solid	→
	→	0.610	0.616	0.619	0.622	0.626	0.627	0.628	0.633	0.643	0.646	→
	→	cc::black	pc::black	tc::white	bkc::black	bll::same-as-head	ts::notched-tail	uptc::black	brc::white	bls::all-purpose	brp::solid	→
	→	0.647	0.651	0.654	0.655	0.659	0.666	0.667	0.669	0.669	0.670	→
	→	bll::shorter-than-head	bec::white	ws::rounded-wings	unc::white	untc::black	upc::black	blc::black	bkp::solid	tp::solid	wc::black	→
	→	0.681	0.681	0.682	0.684	0.685	0.688	0.688	0.691	0.691	0.691	→
	→	sh::perching-like	si::small									
	→	0.693	0.693									

Table 5. **CUB task hardness.** CUB attributes sorted in ascending order of hardness along with their respective hardness scores. Hardness scores listed above are compared with empirical test errors for each task and shown to be strongly correlated (Fig. 6(c) in the paper). Attribute names are abbreviated due to space concerns. Full names are provided in Table 6.

bls	has bill shape	uptc	has upper tail color	untc	has under tail color	tp	has tail pattern
wc	has wing color	hp	has head pattern	nc	has nape color	bep	has belly pattern
upc	has upperparts color	brc	has breast color	bec	has belly color	pc	has primary color
unc	has underparts color	tc	has throat color	ws	has wing shape	lc	has leg color
brp	has breast pattern	ec	has eye color	si	has size	blc	has bill color
bkc	has back color	bll	has bill length	sh	has shape	cc	has crown color
ts	has tail shape	fc	has forehead color	bkp	has back pattern	wp	has wing pattern

Table 6. **CUB attribute name abbreviations.** Abbreviations used in Table 5 for the attributes in the CUB dataset [11].

References

[1] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015. 2

[2] Q. Dong, S. Gong, and X. Zhu. Class rectification hard mining for imbalanced deep learning. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 1851–1860, 2017. 5

[3] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *Trans. Pattern Anal. Mach. Intell.*, 40(11):2597–2609, 2018. 5

[4] E. M. Hand and R. Chellappa. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *AAAI Conf. on Artificial Intelligence*, 2017. 5

[5] C. Huang, Y. Li, C. Change Loy, and X. Tang. Learning deep representation for imbalanced classification. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 5375–5384, 2016. 5

[6] Y. Jang, H. Gunes, and I. Patras. Registration-free face-ssd: Single shot analysis of smiles, facial attributes, and affect in the wild. *Comput. Vision Image Understanding*, 2019. 5

[7] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. Int. Conf. Comput. Vision*, 2015. 3, 5, 9

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learning Research*, 12:2825–2830, 2011. 2

[9] E. M. Rudd, M. Günther, and T. E. Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *European Conf. Comput. Vision*, pages 19–35. Springer, 2016. 5

[10] J. Wang, Y. Cheng, and R. Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 2295–2304, 2016. 5

[11] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 9, 11

[12] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *Trans. Pattern Anal. Mach. Intell.*, 2018. 6, 9

[13] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *Proc. Conf. Comput. Vision Pattern Recognition*, pages 3712–3722, 2018. 2

[14] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *Int. Conf. on Learning Representations*, 2017. 2