Supplementary Material of
Understanding Few-Shot Learning: Measuring Task Relatedness and Adaptation Difficulty via Attributes

1 Theoretical Results and Proofs

Recall that we have defined the distance between two categories y_k, y_t in Section 4.1 of the main paper, expressed as $d(y_k, y_t) = \sum_{a_i \in \mathcal{A}} \frac{1}{2L} \sum_{l=1}^{L} \left|p(a_i^l|y_k) - p(a_i^l|y_t)\right|$ if attribute space $\mathcal{A}$ is countable. Based on the distance definition, we introduce the following theoretical results and proofs.

Lemma 1. *Let $\mathcal{A}$ be the attribute space, L be the number of attributes. Assume all attributes are independent of each other given the class label, i.e. $p(a|y) = \prod_{l=1}^{L} p(a^l|y)$. For all $a_i \in \mathcal{A}$ and any two categories y_k, y_t, the following inequality holds:*

$$\sum_{a_i \in \mathcal{A}} |p(a_i|y_k) - p(a_i|y_t)| \le d(y_k, y_t) + \Delta, \tag{1}$$

where $\Delta = \sum_{a_i \in \mathcal{A}} \frac{1}{2L} \sum_{l=1}^{L} (p(a_i^l|y_k) + p(a_i^l|y_t))$.

Proof. Firstly, the following inequality holds,

$$\frac{1}{L} \sum_{i=1}^{L} x_i \ge (\prod_{i=1}^{L} x_i)^{\frac{1}{L}} \ge \prod_{i=1}^{L} x_i, \tag{2}$$

where x_i is a non-negative real number and ranges from 0 to 1. The proof of Eq. (2) is straightforward: the first inequality is an application of AM-GM inequality (or the inequality of arithmetic and geometric means) on a list of L non-negative real numbers $\{x_1, ..., x_L\}$, and the second inequality holds because $\prod_{i=1}^{L} x_i$ ranges from 0 to 1.

Next, we try to prove Eq. (1) based on the above inequality. If all attributes are independent of each other given the class label, for any conditional probabilities $p(a_i|y_k)$ and $p(a_i|y_t)$, we have

$$\begin{aligned} |p(a_i|y_k) - p(a_i|y_t)| &\le \max(p(a_i|y_k), p(a_i|y_t)) \\ &= \max(\prod_{l=1}^{L} p(a_i^l|y_k), \prod_{l=1}^{L} p(a_i^l|y_t)). \end{aligned} \tag{3}$$

Note that $p(a_i^l|y_k)$ and $p(a_i^l|y_t)$ are both real numbers between 0 and 1. Thus, combining with Eq. (2), we have

$$\begin{aligned} |p(a_i|y_k) - p(a_i|y_t)| &\le \max(\frac{1}{L} \sum_{l=1}^{L} p(a_i^l|y_k), \frac{1}{L} \sum_{l=1}^{L} p(a_i^l|y_t)) \\ &\le \frac{1}{L} \sum_{l=1}^{L} \max(p(a_i^l|y_k), p(a_i^l|y_t)) \\ &\le \frac{1}{2L} \sum_{l=1}^{L} (\left|p(a_i^l|y_k) - p(a_i^l|y_t)\right|) + \frac{1}{2L} \sum_{l=1}^{L} (p(a_i^l|y_k) + p(a_i^l|y_t)). \end{aligned} \tag{4}$$

For all $a_i \in \mathcal{A}$, denote $\Delta = \sum_{a_i \in \mathcal{A}} \frac{1}{2L} \sum_{l=1}^{L} (p(a_i^l|y_k) + p(a_i^l|y_t))$, we have

$$\sum_{a_i \in \mathcal{A}} |p(a_i|y_k) - p(a_i|y_t)| \le d(y_k, y_t) + \Delta. \tag{5}$$

$\square$

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Theorem 1. *With the same notation and assumptions as Lemma 1, let $\mathcal{H}$ be the hypothesis space with VC-dimension d, f_θ and g_ϕ be the meta-learner and the base-learner as introduced in Section 4.2 respectively. Denote g_{ϕ^*} as the best base-learner on some specific task given a fixed meta-learner f_θ. For any training task $\tau_i = (\mathcal{D}_i, S_i)$ and novel task $\tau'_j = (\mathcal{D}'_j, S'_j)$, suppose the number of categories in the two tasks is the same, then with probability at least $1-\delta$, $\forall g_\phi \circ f_\theta \in \mathcal{H}$, we have*

$$\epsilon(f_\theta, \tau'_j) \leq \hat{\epsilon}(f_\theta, \tau_i) + \sqrt{\frac{4}{m_i}\left(d \log \frac{2em_i}{d} + \log \frac{4}{\delta}\right)} + d_\theta(\tau_i, \tau'_j) + \Delta' + \lambda, \tag{6}$$

where $\lambda = \lambda_i + \lambda'_j$ is the generalization error of g_{ϕ^*} and f_θ on the two tasks, i.e., $\lambda_i = \mathbb{E}_{(x,y)\sim\mathcal{D}_i}[\mathbb{I}(g_{\phi_i^*}(f_\theta(x)) \neq y)]$, $\lambda'_j = \mathbb{E}_{(x,y)\sim\mathcal{D}'_j}[\mathbb{I}(g_{\phi'^*_j}(f_\theta(x)) \neq y)]$. Δ' is a term depending on learned base-learners $g_{\phi_i}, g_{\phi'_j}$ and the best base-learners $g_{\phi_i^*}, g_{\phi'^*_j}$.

Proof. Note that the error $\epsilon(f_\theta, \tau'_j) = \mathbb{E}_{(x,y)\sim\mathcal{D}'_j}[\mathbb{I}(g_{\phi'_j}(f_\theta(x)) \neq y)]$ can be decomposed into two parts: with the same f_θ, (1) the probability that the learned base-learner $g_{\phi'_j}$ agrees with the best base-learner $g_{\phi_j^*}$, but they both output the wrong prediction; and (2) the probability that the learned base-learner $g_{\phi'_j}$ disagrees with the best base-learner $g_{\phi_j^*}$, while $g_{\phi'_j}$ outputs the wrong prediction. The first part can be bounded by the error of $g_{\phi_j^*}$, and the second part can be bounded by the probability that $g_{\phi'_j}$ disagrees with $g_{\phi_j^*}$. Denote $Z^{j'} = \{(x,y) | g_{\phi'_j}(f_\theta(x)) \neq g_{\phi_j^*}(f_\theta(x)), (x,y) \sim \mathcal{D}'_j\}$, then $P_{\mathcal{D}'_j}[Z^{j'}]$ represents the probability that $g_{\phi'_j}$ disagrees with $g_{\phi_j^*}$ based on the same f_θ on distribution $\mathcal{D}'_j$. We have

$$\begin{aligned}
\epsilon(f_\theta, \tau'_j) &\leq \lambda'_j + P_{\mathcal{D}'_j}[Z^{j'}] \\
&= \lambda'_j + P_{\mathcal{D}_i}[Z^i] + P_{D'_j}[Z^{j'}] - P_{\mathcal{D}_i}[Z^i] \\
&\leq \lambda'_j + \lambda_i + \epsilon(f_\theta, \tau_i) + P_{\mathcal{D}'_j}[Z^{j'}] - P_{D_i}[Z^i] \\
&= \lambda + \epsilon(f_\theta, \tau_i) + P_{\mathcal{D}'_j}[Z^{j'}] - P_{\mathcal{D}_i}[Z^i].
\end{aligned} \tag{7}$$

Assume that the two tasks both have C categories and $p(y)$ is uniform, we can decompose the probability $P_{\mathcal{D}'_j}[Z^{j'}]$ as $P_{\mathcal{D}'_j}[Z^{j'}] = \frac{1}{C}\sum_{t=1}^{C} P_{x|y_t}[Z_t^{j'}]$, where $Z_t^{j'} = \{x | g_{\phi'_j}(f_\theta(x)) \neq g_{\phi_j^*}(f_\theta(x)), x \sim p(x|y_t)\}$. Thus, we have

$$\begin{aligned}
\epsilon(f_\theta, \tau'_j) &\leq \lambda + \epsilon(f_\theta, \tau_i) + \frac{1}{C}\left(\sum_{t=1}^{C} P_{x|y_t}[Z_t^{j'}] - \sum_{k=1}^{C} P_{x|y_k}[Z_k^i]\right) \\
&= \lambda + \epsilon(f_\theta, \tau_i) + \frac{1}{C}\sum_{e_{tk}\in M}\left(P_{x|y_t}[Z_t^{j'}] - P_{x|y_k}[Z_k^i]\right),
\end{aligned} \tag{8}$$

where M is a maximum matching, which contains C edges and each edge $e_{tk} \in M$ links two categories y_t, y_k in task τ'_j and τ_i respectively.

Next, we consider to replace the conditional distribution $p(x|y)$ with the attribute conditional distribution $p(a|y)$, because the former is usually unknown and difficult to estimate. For a conditional distribution $p(x|y)$ and a mapping $f_\theta : \mathcal{X} \to \mathcal{A}$, a new distribution can be induced over the space $\mathcal{A}$ as $p_\theta(a|y) \triangleq p(f_\theta(x)|y)$. Based on the induced distribution $p_\theta(a|y)$, we have $P_{x|y_t}[Z_t^{j'}] = P_{a|y_t}[\{a | g_{\phi'_j}(a) \neq g_{\phi_j^*}(a), a \sim p_\theta(a|y_t)\}]$. For clarity, we define $A_t = \{a | g_{\phi'_j}(a) \neq g_{\phi_j^*}(a), a \sim p_\theta(a|y_t)\}$ and $A_k = \{a | g_{\phi_i}(a) \neq g_{\phi_i^*}(a), a \sim p_\theta(a|y_k)\}$. Thus, Eq. (8) can be rewritten as

$$\epsilon(f_\theta, \tau'_j) \leq \lambda + \epsilon(f_\theta, \tau_i) + \frac{1}{C}\sum_{e_{tk}\in M}\left(P_{a|y_t}[A_t] - P_{a|y_k}[A_k]\right). \tag{9}$$

Let $A_{t\cup k} = A_t \cup A_k$ be the union set of A_t and A_k, $A_{t\cap k} = A_t \cap A_k$ be the intersection set of A_t and A_k, then we have another inequality as

$$\begin{aligned}
P_{a|y_t}[A_t] - P_{a|y_k}[A_k] \leq &\left(P_{a|y_t}[A_{t\cup k}] - P_{a|y_k}[A_{t\cup k}]\right) + \left|P_{a|y_t}[A_{t\cap k}] - P_{a|y_k}[A_{t\cap k}]\right| \\
&+ \left|P_{a|y_t}[A_k] - P_{a|y_k}[A_t]\right|.
\end{aligned} \tag{10}$$

For clarity, we use two notions Δ_1 and Δ_2 to denote $\frac{1}{C}\sum_{e_{tk}\in M}\left|P_{a|y_t}[A_{t\cap k}]-P_{a|y_k}[A_{t\cap k}]\right|$ and $\frac{1}{C}\sum_{e_{tk}\in M}\left|P_{a|y_t}[A_k]-P_{a|y_k}[A_t]\right|$, respectively. Based on Lemma 1 and Eq. (10), we have

$$\begin{aligned}\frac{1}{C}\sum_{e_{tk}\in M}(P_{a|y_t}[A_t]-P_{a|y_k}[A_k]) &\le \frac{1}{C}\sum_{e_{tk}\in M}(P_{a|y_t}[A_{t\cup k}]-P_{a|y_k}[A_{t\cup k}])+\Delta_1+\Delta_2\\ &=\frac{1}{C}\sum_{e_{tk}\in M}\sum_{a_i\in A_{t\cup k}}(p_\theta(a_i|y_t)-p_\theta(a_i|y_k))+\Delta_1+\Delta_2\\ &\le \frac{1}{C}\sum_{e_{tk}\in M}d_\theta(y_t,y_k)+\Delta+\Delta_1+\Delta_2\\ &=d_\theta(\tau'_j,\tau_i)+\Delta+\Delta_1+\Delta_2,\end{aligned}\tag{11}$$

where $\Delta=\frac{1}{C}\sum_{e_{tk}\in M}\sum_{a_i\in A_{t\cup k}}\frac{1}{2L}\sum_{l=1}^{L}(p_\theta(a_i^l|y_k)+p_\theta(a_i^l|y_t))$. Denoting $\Delta'=\Delta+\Delta_1+\Delta_2$, and combining Eq. (9) and Eq. (11), we can get

$$\epsilon(f_\theta,\tau'_j)\le\lambda+\epsilon(f_\theta,\tau_i)+d_\theta(\tau_i,\tau'_j)+\Delta'.\tag{12}$$

Finally, we apply Vanik-Chervonenkis theory [11] to bound the generalization error $\epsilon(f_\theta,\tau_i)$ in Eq. (12) by its empirical estimate $\hat{\epsilon}(f_\theta,\tau_i)$. Namely, if S_i is a m_i-size i.i.d sample set, then with probability at least $1-\delta$,

$$\epsilon(f_\theta,\tau_i)\le\hat{\epsilon}(f_\theta,\tau_i)+\sqrt{\frac{4}{m_i}(d\log\frac{2em_i}{d}+\log\frac{4}{\delta})}.\tag{13}$$

Combining with Eq. (12), with probability at least $1-\delta$, we have

$$\epsilon(f_\theta,\tau'_j)\le\hat{\epsilon}(f_\theta,\tau_i)+\sqrt{\frac{4}{m_i}(d\log\frac{2em_i}{d}+\log\frac{4}{\delta})}+d_\theta(\tau_i,\tau'_j)+\Delta'+\lambda.\tag{14}$$

□

Corollary 1. *With the same notation and assumptions as Theorem 1, for* n *training tasks* $\{\tau_i\}_{i=1}^n$ *and a novel task* τ'_j*, define* $\hat{\epsilon}(f_\theta,\tau_{i=1}^n)=\frac{1}{n}\sum_{i=1}^n\hat{\epsilon}(f_\theta,\tau_i)$*, then with probability at least* $1-\delta$*,* $\forall g_\phi\circ f_\theta\in\mathcal{H}$*, we have*

$$\epsilon(f_\theta,\tau'_j)\le\hat{\epsilon}(f_\theta,\tau_{i=1}^n)+\frac{1}{n}\sum_{i=1}^n\sqrt{\frac{4}{m_i}(d\log\frac{2em_i}{d}+\log\frac{4}{\delta})}+\frac{1}{n}\sum_{i=1}^n d_\theta(\tau_i,\tau'_j)+\Delta'+\lambda,\tag{15}$$

where $\lambda=\frac{1}{n}\sum_{i=1}^n\lambda_i+\lambda'_j$*, and* Δ' *is a term depending on the learned base-learners* $\{g_{\phi_i}\}_{i=1}^n, g_{\phi'_j}$ *and the best base-learners* $\{g_{\phi_i^*}\}_{i=1}^n, g_{\phi_j'^*}$*.*

Proof. The proof of Corollary 1 is similar to the proof of Theorem 1. Denote $\lambda=\frac{1}{n}\sum_{i=1}^n\lambda_i+\lambda'_j$, we have

$$\begin{aligned}\epsilon(f_\theta,\tau'_j)&\le\lambda'_j+P_{\mathcal{D}'_j}[Z^{j'}]\\ &=\lambda'_j+\frac{1}{n}\sum_{i=1}^n P_{\mathcal{D}_i}[Z^i]+P_{\mathcal{D}'_j}[Z^{j'}]-\frac{1}{n}\sum_{i=1}^n P_{\mathcal{D}_i}[Z^i]\\ &\le\lambda'_j+\frac{1}{n}\sum_{i=1}^n\lambda_i+\epsilon(f_\theta,\tau_{i=1}^n)+P_{\mathcal{D}'_j}[Z^{j'}]-\frac{1}{n}\sum_{i=1}^n P_{D_i}[Z^i]\\ &=\lambda+\epsilon(f_\theta,\tau_{i=1}^n)+\frac{1}{n}\sum_{i=1}^n(P_{\mathcal{D}'_j}[Z^{j'}]-P_{\mathcal{D}_i}[Z^i]).\end{aligned}\tag{16}$$

Now, we can follow the same procedure as the proof in Theorem 1 and have the following inequlaity

$$\epsilon(f_\theta,\tau'_j)\le\lambda+\hat{\epsilon}(f_\theta,\tau_{i=1}^n)+\frac{1}{n}\sum_{i=1}^n\sqrt{\frac{4}{m_i}(d\log\frac{2em_i}{d}+\log\frac{4}{\delta})}+\frac{1}{n}\sum_{i=1}^n d_\theta(\tau_i,\tau'_j)+\Delta',\tag{17}$$

where $\Delta'=\frac{1}{n}\sum_{i=1}^n\Delta'_i$, and Δ'_i corresponds to the additional non-negative term as in Eq. (12), which is derived from $P_{\mathcal{D}'_j}[Z^{j'}]-P_{\mathcal{D}_i}[Z^i]$. □

Theorem 2. *With the same notation and assumptions as in Corollary 1, assume the conditional distribution $p(x|a^l)$ is task agnostic. If the number of labeled samples m_i in n training tasks and the number of labeled samples m'_j in novel task τ'_j tend to be infinite, the following inequality holds:*

$$\frac{1}{n}\sum_{i=1}^{n} d_\theta(\tau_i, \tau'_j) \leq \frac{1}{n}\sum_{i=1}^{n} d(\tau_i, \tau'_j). \tag{18}$$

Proof. Without loss of generality, we first consider a single training task τ_i and prove $d_\theta(\tau_i, \tau'_j) \leq d(\tau_i, \tau'_j)$. Assume the number of training samples is infinite. Thus, for any category y_k in training task τ_i, the induced distribution $p^i_\theta(a|y_k)$ is equal to the ground-truth distribution $p^i(a|y_k)$. However, for any category y_t in novel task τ'_j, even with the infinite novel samples, the induced distribution $p^j_\theta(a|y_t)$ does not equal the ground-truth distribution $p^j(a|y_t)$. This is because we fix f_θ and train a new base-learner $g_{\phi'_j}$ to adapt to novel task τ'_j, as introduced in Section 4.2. Thus, we have $d_\theta(\tau_i, \tau'_j) = \frac{1}{C}\sum_{e_{kt}\in M}\frac{1}{L}\sum_{l=1}^{L} d_{L_1}(p^i(a^l|y_k), p^j_\theta(a^l|y_t))$, which measures the distance between the ground-truth distribution $p^i(a^l|y_k)$ and the induced distribution $p^j_\theta(a^l|y_t)$. Next, for any attribute l, we consider three cases: (1) the values of attribute l in training and novel tasks are disjoint, which means it is a new attribute or new values of observed attribute for novel task τ'_j. In this case, for any two categories y_k and y_t, the model-related distance $d_{L_1}(p^i(a^l|y_k), p^j_\theta(a^l|y_t)) \leq d_{L_1}(p^i(a^l|y_k), p^j(a^l|y_t)) = 1$; (2) the values of attribute l in training and novel tasks are completely overlapped. As the conditional distribution $p(x|a^l)$ is task-agnostic, the attribute classifier f_θ can also identify attribute l in novel task τ'_j. In this case, $d_{L_1}(p^i(a^l|y_k), p^j_\theta(a^l|y_t)) = d_{L_1}(p^i(a^l|y_k), p^j(a^l|y_t))$; (3) the values of attribute l in training and novel tasks are overlapped but not the same. We can divide the values of attribute l into two parts: the completely overlapped values and the disjoint values, then we can follow the same analysis procedures as in the case (1) and case (2).

In summary, we can arrive that the model-related distance $d_\theta(\tau_i, \tau'_j)$ is no more than the model-agnostic distance $d(\tau_i, \tau'_j)$, thus we have $\frac{1}{n}\sum_{i=1}^{n} d_\theta(\tau_i, \tau'_j) \leq \frac{1}{n}\sum_{i=1}^{n} d(\tau_i, \tau'_j)$.

□

2 Attribute Prototypical Network

Our theoretical analysis is based on a specific meta-learning framework with attribute learning. Thus, we instantiate a simple model under that framework as an example, we call this model Attribute Prototypical Network (APNet). A sketch of APNet is presented in Fig. 1.

Let $S = \{(x_k, y_k)\}_{k=1}^{m}$ include all labeled samples in n training tasks. For each sample $(x_k, y_k) \in S$, assume we have L binary attribute labels $\{a^l_k\}_{l=1}^{L}$. As Corollary 1 reveals, we can reduce the generalization error on novel tasks by maximizing the attribute discrimination ability of meta-learner f_θ and the classification ability of base-learner g_ϕ. Specifically, we adopt a convolutional network with an additional MLP as f_θ. The convolutional network extracts feature representations from images, then the MLP takes features as input and output attribute labels. The attribute classification loss is defined as

$$\mathcal{L}(f_\theta) = -\frac{1}{m}\sum_{k=1}^{m}\frac{1}{L}\sum_{l=1}^{L}\left[a^l_k \log z^l_k + (1-a^l_k)\log(1-z^l_k)\right], \tag{19}$$

where z^l_k is the l-th dimension of $f_\theta(x_k)$ after a sigmoid function.

For base-learner, we simply choose an non-parametric base-learner like ProtoNet [9][1], which takes the attributes generated by the meta-learner f_θ as input to calculate cosine distance between test samples and attribute prototypes, then predicts the target label. The few-shot classification loss is

---

[1] Any other models that map $a \in \mathcal{A}$ to $y \in \mathcal{Y}$ are also feasible, such as a MLP in Relation Network [10] and a parametric cosine classifier in Baseline++ [3].

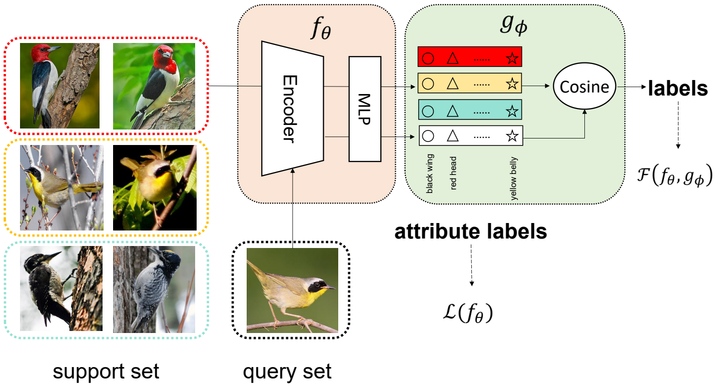

Figure 1: A sketch of APNet.

defined as

$$\mathcal{F}(f_\theta, g_{\phi_i}) = -\frac{1}{m_i}\sum_{k=1}^{m_i} y_k \log \frac{exp(d_k/t)}{\sum_{y_k} exp(d_k/t)}, \tag{20}$$

where $d_k \triangleq cos(f_\theta(x_k), c_{y_k})$ denotes cosine similarity and c_{y_k} denotes the attribute prototype of category y_k. t is a scalar temperature factor. In practice, we use a hyperparameter β to balance the two losses, so that the final training objective is

$$\mathcal{L} = \beta * \mathcal{L}(f_\theta) + \frac{1}{n}\sum_{i=1}^{n} \mathcal{F}(f_\theta, g_{\phi_i}). \tag{21}$$

During the inference phrase, we fix f_θ then calculate the cosine distance between each query sample and attribute prototypes to predict the target label.

3 Experiment Details

3.1 Implementation Details

We run experiments with APNet and five classical FSL methods (MatchingNet [12], ProtoNet [9], RelationNet [10], MAML [4], Baseline++ [3]). Here we explain more implementation details about these methods. As existing work [3] has provided a unified testbed for several different FSL methods, we use the codebase and run the experiments for the above methods. For a fair comparison, we use the four-layer convolution network (Conv4) as backbone model for all methods. On the CUB dataset, we perform standard data augmentation, including random crop, rotation, horizontal flipping and color jittering, as in [2]. On the SUN dataset, we simply use two augmentation operations, including image scaling and horizontal flipping. For APNet, we use all provided attribute information (attribute locations and labels) to calculate the attribute classification loss $\mathcal{L}(f_\theta)$. Because the SUN dataset does not provide attribute locations, we only use attribute labels to calculate $\mathcal{L}(f_\theta)$. We use the Adam optimizer [5] with an initial learning rate of 10^{-3} and weight decay of 0. We train models on 5-shot tasks for 40,000 episodes and on 1-shot tasks for 60,000 episodes. The hyperparameter β is tuned on the validation set. We set β to 0.6 and 1.0 for 1-shot and 5-shot setting respectively on CUB dataset, and 0.6 for both settings on SUN dataset.

3.2 Complete Results

Here we show complete experimental results which have been partially shown in the main paper. Tab. 1 shows the results on CUB and SUN dataset. Tab. 2 shows the results on *mini*ImageNet and the cross-dataset scenario (*mini*ImageNet → CUB).

Table 1: 5-way 1-shot and 5-shot performance of different FSL methods on CUB and SUN datasets. Conv4 is used as the backbone model. We report the average accuracy on 600 novel tasks with 95% confidence interval.

Method	**Backbone**	**CUB**		**SUN**	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
MatchingNet	Conv4	61.02 (0.88)	79.99 (0.75)	57.87 (0.95)	76.80 (0.68)
ProtoNet		57.12 (0.94)	76.67 (0.65)	60.20 (0.90)	76.75 (0.65)
RelationNet		61.86 (0.98)	76.63 (0.71)	60.52 (0.91)	76.49 (0.65)
MAML		58.73 (0.97)	76.20 (0.69)	59.65 (0.94)	76.82 (0.68)
Baseline++		60.57 (0.80)	80.17 (0.61)	49.78 (0.82)	74.09 (1.11)
APNet	Conv4	72.96 (0.89)	85.48 (0.55)	60.53 (0.86)	76.35 (0.63)

Table 2: 5-way 1-shot and 5-shot performance of different FSL methods on *mini*ImageNet and cross-dataset scenario (*mini*ImageNet→CUB). Conv4 is used as the backbone model. We report the average accuracy on 600 novel tasks with 95% confidence interval.

Method	**Backbone**	***mini*ImageNet**		***mini*ImageNet→CUB**	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
MatchingNet	Conv4	49.36 (0.79)	62.77 (0.69)	37.48 (0.68)	49.98 (0.66)
ProtoNet		42.53 (0.84)	62.89 (0.72)	33.91 (0.67)	53.74 (0.72)
RelationNet		48.38 (0.80)	64.37 (0.72)	38.19 (0.69)	52.57 (0.66)
MAML		45.70 (0.85)	62.64 (0.72)	36.97 (0.69)	51.60 (0.70)
Baseline++		47.01 (0.71)	66.72 (0.62)	37.11 (0.66)	52.42 (0.67)

4 Additional Experiments

4.1 Deeper Backbone

As mentioned in the main paper, we have shown that TAD can serve as a metric to measure the adaptation difficulty on novel tasks for different FSL methods. Here we consider how a deeper backbone affects this conclusion. Following [3], we use ResNet18 as backbone model and train the five FSL models. The experimental results are shown in Tab. 3. Fig. 2 shows the task distance and the corresponding accuracy of 2,400 novel tasks. As shown in Fig. 2, we observe similar phenomenon that with the increase of task distance, the accuracy of these models tends to decrease. This indicates that the proposed TAD metric works for different FSL methods with a deeper backbone model.

4.2 Comparison with Other Metrics

Here we present comparisons between proposed TAD and other metrics to show the effectiveness of it. For comparing different metrics, we design a task selection experiment. More specifically, we select top 5% novel tasks with the highest distances computed by different metrics, and then evaluate the accuracy of FSL models on these chosen tasks. The central hypothesis behind this experiment is

Table 3: 5-way 1-shot and 5-shot performance of different FSL methods on CUB and SUN. ResNet18 is used as the backbone model. We report the average accuracy on 600 novel tasks with 95% confidence interval.

Method	**Backbone**	**CUB**		**SUN**	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
MatchingNet	ResNet18	74.62 (0.87)	85.02 (0.54)	67.42 (0.90)	78.30 (0.67)
ProtoNet		74.04 (0.88)	87.30 (0.49)	67.71 (0.87)	81.68 (0.60)
RelationNet		70.37 (0.98)	84.41 (0.57)	63.85 (0.93)	79.60 (0.67)
MAML		70.76 (1.04)	81.31 (0.70)	61.47 (0.98)	75.24 (0.73)
Baseline++		70.20 (0.93)	84.11 (0.57)	53.06 (0.77)	74.21 (0.68)

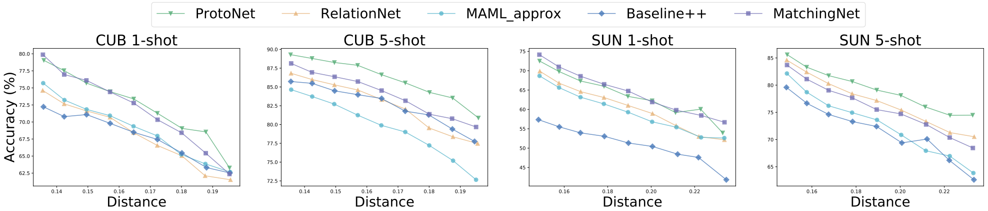

Figure 2: Accuracy of different methods in terms of the average task distance. From left to right, 5-way 1-shot and 5-shot on CUB/SUN. ResNet18 is used as the backbone model.

Table 4: The difference in accuracy between the chosen tasks and all novel tasks for various FSL models. In the last column, we report the computation time (seconds) of different metrics. We run the experiment on the CUB dataset with 5-way 1-shot setting.

Metrics	MatchingNet	ProtoNet	RelationNet	Baseline++	APNet	Time
FID	-3.71	-1.78	-0.57	-0.33	-3.63	480
EMD	-0.24	-1.33	1.60	-0.75	1.17	22
Task2Vec	-0.54	-0.30	-3.64	-0.54	-2.45	6000
TAD	**-8.23**	**-6.64**	**-7.39**	**-3.52**	**-7.24**	**3**

that if a distance metric can better reflect task difficulty, then novel tasks with the highest distances should be more challenging. We choose three methods for comparison, which have been proposed in the few-shot learning or related area: (1) Frechet Inception Distance (**FID**) [6], FID is a metric to measure the distance between two image distributions by comparing their mean and covariance. (2) Earth Mover's Distance (**EMD**) [7], EMD is a measure of dissimilarity between two distributions by considering the distance as the cost of moving images from one distribution to the other. (3) **Task2Vec** [1], TaskVec is a task embedding method which represents each task as an embedding with Fisher Information Matrix, and the norm of embedding reflects task difficulty. Note that the above three methods rely on a pretrained model. Following [7], we use ResNet-101 pre-trained on ImageNet for them. Tab. 4 illustrates the results of different metrics on the CUB dataset. We find that, with human-annotated attributes, TAD significantly outperforms other three methods in identifying more challenging novel tasks across all FSL models, demonstrating the effectiveness of TAD metric. Furthermore, the computational efficiency of TAD greatly surpasses other methods, as illustrated in the last column of table. Notably, TAD requires only 3 seconds to compute across 2400 novel tasks, underscoring its advantage of ease of computation.

5 Analysis of Auto-Annotated Attributes

We try to evaluate the quality of the auto-annotated attributes generated by pretrained CLIP and then give some examples for qualitative analysis. Due to the absence of attribute annotations in the miniImagenet dataset, we collect the annotations ourselves. Initially, we predefine 25 attribute labels (as shown in Tab. 5) following [8], and then randomly select 50 images for annotation. By comparing the ground-truth annotations with the results produced by the CLIP model, we discern that the average accuracy across the 25 attributes approaches 0.65. Notably, we observe that CLIP achieves good performance across the majority of attributes, with accuracies ranging from 0.7 to 0.9. However, it fails on some attributes such as "white," "pink," "smooth," and "shiny," where the accuracy decreases to approximately 0.2. Fig. 3 shows qualitative examples of the obtained attributes. We find that the common wrong cases are color attributes, as the CLIP model always predicts more colors than manual annotations. Thus, in all other experiments under the cross-dataset scenario, we remove 11 color attributes and only considere the remaining 14 attributes when calculating the task distance for the TAD metric.

References

[1] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-

Table 5: Details of pre-defined 25 attribute labels, including color, pattern, shape and texture.

	Attributes
Color	black, blue, brown, gray, green, orange, pink, red, violet, white, yellow
Pattern	spotted, striped
Shape	long, round, rectangular, square
Texture	furry, smooth, rough, shiny, metallic, vegetation, wooden, wet

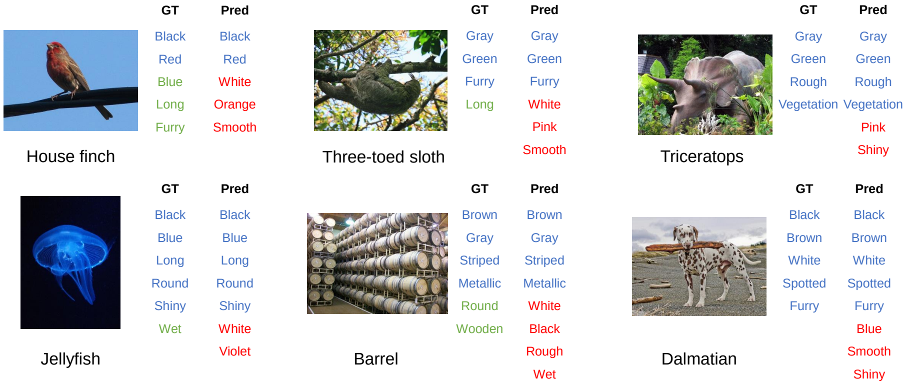

Figure 3: Qualitative examples of auto-annotated attributes. Attributes predicted by the CLIP model and overlapped with the ground truth are highlighted in blue, while the ground truth without match is marked in green. False positive predictions are represented in red.

learning. In *ICCV*, 2019.

[2] Kaidi Cao, Maria Brbic, and Jure Leskovec. Concept learners for few-shot learning. In *ICLR*, 2021.

[3] Weiyu Chen, Yencheng Liu, Zsolt Kira, Yuchiang Frank Wang, and Jiabin Huang. A closer look at few-shot classification. In *ICLR*, 2019.

[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[6] Timo Milbich, Karsten Roth, Samarth Sinha, Ludwig Schmidt, Marzyeh Ghassemi, and Bjorn Ommer. Characterizing generalization under out-of-distribution shifts in deep metric learning. In *NeurIPS*, 2021.

[7] Jaehoon Oh, Sungnyun Kim, Namgyu Ho, Jin-Hwa Kim, Hwanjun Song, and Se-Young Yun. Understanding cross-domain few-shot learning based on domain similarity and few-shot difficulty. In *NeurIPS*, 2022.

[8] Olga Russakovsky and Li Fei-Fei. Attribute learning in large-scale datasets. In *ECCV*, 2010.

[9] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[10] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018.

[11] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[12] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, 2016.