Supplementary Material

A. Proof

This appendix collects all the proofs omitted from the main text.

A.1. Proof of Theorem 1

This section provides a detailed proof for Theorem 1, which is ommitted from the main text. We first recall two lemmas by Bartlett et al. [32] .

Lemma 3 (cf. [32], Lemma A.7)**.** *Suppose there are L weight matrices in a chain-like neural network. Let $(\varepsilon_1, \dots, \varepsilon_L)$ be given. Suppose the L weight matrices $(A_1, \dots, A_L)$ lies in $\mathcal{B}_1 \times \dots \times \mathcal{B}_L$, where $\mathcal{B}_i$ is a ball centered at 0 with the radius of s_i, i.e., $\mathcal{B}_i = \{A_i : \|A_i\| \le s_i\}$. Furthermore, suppose the input data matrix X is restricted in a ball centred at 0 with the radius of B, i.e., $\|X\| \le B$. Suppose F is a hypothesis function computed by the neural network. If we define:*

$$\mathcal{H} = \{F(X) : A_i \in \mathcal{B}_i, A_t^{u,v,s} \in \mathcal{B}_t^{u,v,s}\} , \tag{A.1}$$

where $i = 1, \dots, L$, $(u, v, s) \in I_V$, and $t \in \{1, \dots, L^{u,v,s}\}$. Let $\varepsilon = \sum_{j=1}^{L} \varepsilon_j \rho_j \prod_{l=j+1}^{L} \rho_l s_l$. Then we have the following inequality:

$$\mathcal{N}(\mathcal{H}) \le \prod_{i=1}^{L} \sup_{\mathbf{A}_{i-1} \in \boldsymbol{\mathcal{B}}_{i-1}} \mathcal{N}_i , \tag{A.2}$$

where $\mathbf{A}_{i-1} = (A_1, \dots, A_{i-1})$, $\boldsymbol{\mathcal{B}}_{i-1} = \mathcal{B}_1 \times \dots \times \mathcal{B}_{i-1}$, and

$$\mathcal{N}_i = \mathcal{N}\left(\left\{A_i F_{\mathbf{A}_{i-1}}(X) : A_i \in \mathcal{B}_i\right\} \varepsilon_i, \|\cdot\|\right) . \tag{A.3}$$

Here, the radius of each covers are respectively,

$$\varepsilon_i = \frac{\alpha_i \varepsilon}{\rho_i \prod_{j>i} \rho_j s_j}, \tag{A.4}$$

where

$$\alpha_i = \frac{1}{\bar{\alpha}} \left(\frac{b_i}{s_i}\right)^{2/3}, \tag{A.5}$$

$$\bar{\alpha} = \sum_{j=1}^{L} \left(\frac{b_j}{s_j}\right)^{2/3} . \tag{A.6}$$

This lemma gives an upper bound on the covering number of the hypothesis space of any chain-like neural network.

Lemma 4 (cf. [32], Lemma 3.2)**.** *Let conjugate exponents (p, q) and (r, s) be given with $p \le 2$, as well as positive reals (a, b, ε) and positive integer m. Let matrix $X \in \mathbb{R}^{n \times d}$ be given with $\|X\|_p \le b$. Let $\mathcal{H}_A$ denote the family of matrices obtained by evaluating X with all choices of matrix A:*

$$\mathcal{H}_A \triangleq \left\{XA | A \in \mathbb{R}^{d \times m}, \|A\|_{q,s} \le a\right\} . \tag{A.7}$$

Then

$$\log \mathcal{N}\left(\mathcal{H}_A, \varepsilon, \|\cdot\|_2\right) \le \left\lceil \frac{a^2 b^2 m^{2/r}}{\varepsilon^2} \right\rceil \log(2dm) . \tag{A.8}$$

This lemma gives an upper bound on the covering number of the hypothesis space contributed by a single weight matrix. Based on Lemmas 3 and 4, we can further obtain the proof for Theorem 1.

Proof of Theorem 1. Suppose the hypothesis spaces of the output functions $F_{(A_1,\dots,A_i)}$ of the weight matrices $A_i, i = 1,\dots,8$ are respectively $\mathcal{H}_i, i = 1,\dots,8$, and the corresponding covering number is $\mathcal{N}_i$. From Lemma 3, we can directly get the following inequality,

$$\begin{aligned}
&\log \mathcal{N}(\mathcal{H}) \\
&\leq \log \left(\prod_{i=1}^{8} \sup_{\mathbf{A}_i \in \boldsymbol{\mathcal{B}}_{i-1}} \mathcal{N}_i \right) \\
&\leq \sum_{i=1}^{8} \log \left(\sup_{\substack{(A_1,\dots,A_i) \\ \forall j<i, A_j \in B_j}} \mathcal{N}\left(\left\{ A_i F_{(A_1,\dots,A_i)} \right\}, \varepsilon_i, \|\cdot\|_2 \right) \right) .
\end{aligned} \tag{A.9}$$

Employ eq. (A.8), we can get the following inequality,

$$\log \mathcal{N}(\mathcal{H}) \leq \sum_{i=1}^{8} \frac{b_i^2 \|F_{(A_1,\dots,A_i)}(X)\|_\sigma^2}{\varepsilon_i^2} \log\left(2W^2\right) . \tag{A.10}$$

Meanwhile,

$$\begin{aligned}
\|F_{(A_1,\dots,A_i)}(X)\|_\sigma =& \|\sigma_i(A_i F_{(A_1,\dots,A_{i-1})}(X)) - \sigma_i(0)\|_2 \\
\leq& \|\sigma_i\| \|A_i F_{(A_1,\dots,A_{i-1})}(X) - 0\|_2 \\
\leq& \rho_i \|A_i\|_\sigma \|F_{(A_1,\dots,A_{i-1})}(X)\|_2 \\
\leq& \rho_i s_i \|F_{(A_1,\dots,A_{i-1})}(X)\|_2 .
\end{aligned} \tag{A.11}$$

Therefore,

$$\|F_{(A_1,\dots,A_i)}(X)\|_\sigma^2 \leq \|X\|^2 \prod_{j=1}^{i} s_j^2 \rho_j^2 . \tag{A.12}$$

Applying eqs. (A.5) and (A.6),

$$\begin{aligned}
\log \mathcal{N}(\mathcal{H}) \leq& \sum_{i=1}^{8} \frac{b_i^2 \|X\|^2 \prod_{j=1}^{i} s_j^2 \rho_j^2}{\varepsilon_i^2} \log\left(2W^2\right) \\
=& \sum_{i=1}^{8} \frac{b_i^2 \|X\|^2 s_8^2 \prod_{j=1}^{7} s_j^2 \rho_j^2}{\varepsilon_i^2} \log\left(2W^2\right) \sum_{i=1}^{8} \frac{b_i^2}{\alpha_i^2 s_i^2} \\
=& \sum_{i=1}^{8} \frac{b_i^2 \|X\|^2 s_8^2 \prod_{j=1}^{7} s_j^2 \rho_j^2}{\varepsilon_i^2} \log\left(2W^2\right) \left(\bar{\alpha}^3\right) \\
=& \sum_{i=1}^{8} \frac{b_i^2 \|X\|^2 s_8^2 \prod_{j=1}^{7} s_j^2 \rho_j^2}{\varepsilon_i^2} \log\left(2W^2\right) \left(\sum_{j=1}^{8} \frac{b_j^{2/3}}{s_j^{2/3}} \right)^3 .
\end{aligned} \tag{A.13}$$

The proof is completed. □

A.2. Proof of Theorem 2

This section provides a detailed proof for Theorem 2, which is ommitted from the main text. We first recall a classic result in learning theory which expresses the negative correlation between the generalization error of an algorithm and the corresponding Rademacher complexity $\hat{\mathfrak{R}}(\mathcal{H})$ as the following lemma.

Lemma 5 (cf. [25], Theorem 3.1)**.** *For any $\delta > 0$, with probability at least $1 - \delta$, the following inequality hold for all $F_\theta \in \mathcal{H}$:*

$$\mathcal{R}(F_\theta) \leq \hat{\mathcal{R}}(F_\theta) + 2\hat{\mathfrak{R}}(l \circ \mathcal{H}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2N}} , \tag{A.14}$$

where $l \circ \mathcal{H}$ is defined as

$$l \circ \mathcal{H} \triangleq \{l \circ F : \, F \in \mathcal{H}\} \, . \tag{A.15}$$

Computing the empirical Rademacher complexity of neural network could be extremely difficult and thus still remains an open problem. Fortunately, the empirical Rademacher complexity can be upper bounded by the corresponding ε-covering number $N(\mathcal{H}, \varepsilon, \| \cdot \|_2)$ as the following lemma states.

Lemma 6 (cf. [32], Lemma A.5)**.** *Suppose* $\mathbf{0} \in \mathcal{H}$*, then*

$$\mathfrak{R}(\mathcal{H}) \leq \inf_{\alpha > 0} \left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \int_{\alpha}^{\sqrt{n}} \sqrt{\log \mathcal{N}(l \circ \mathcal{H}, \varepsilon, \| \cdot \|_2)} d\varepsilon \right) \, . \tag{A.16}$$

Proof of Theorem 2. Apply Lemma 6 directly to Theorem 1, we can get the following equation

$$\begin{aligned}\mathfrak{R}(\mathcal{H}) &\leq \inf_{\alpha > 0} \left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \int_{\alpha}^{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{H}_{\lambda}|_D, \varepsilon, \| \cdot |_2)} \mathrm{d}\varepsilon \right) \\ &\leq \inf_{\alpha > 0} \left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \int_{\alpha}^{\sqrt{n}} \frac{R}{\varepsilon} \mathrm{d}\varepsilon \right) \\ &\leq \inf_{\alpha > 0} \left[\frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \sqrt{R} \log \left(\frac{\sqrt{n}}{\alpha} \right) \right] \, . \end{aligned} \tag{A.17}$$

Apparently, the infinimum is reached uniquely at $\alpha = 3\sqrt{\frac{R}{n}}$ and the infinitum is as follows,

$$\mathfrak{R}(\mathcal{H}) \leq \frac{12R}{N} \left[1 + \log \left(\frac{N}{3R} \right) \right] \, . \tag{A.18}$$

Apply eq. (A.18) to eq. (A.14) of Lemma 5, we can directly get the following equation,

$$\mathcal{R}(F_\theta) \leq \hat{\mathcal{R}}(F_\theta) + \frac{24R}{N} \left[1 + \log \left(\frac{N}{3R} \right) \right] + 3\sqrt{\frac{\log \frac{2}{\delta}}{2N}} \, , \tag{A.19}$$

which is exactly eq. (4.6).

The proof is completed. □

B. Empirical Results

This appendix collects all empirical results omitted from the main text. Specifically, we provides some exmaples from the datasets, an example implementation of SAML based on the cosine metric, additional illustrations of semantically relevant local regions, studies on three other metrics, and comparison experiments with the state-of-the-art methods based on deeper networks.

B.1. Examples from the Datasets

This section provides exemplary images from the two datasets used in the experiments as Figure 7.

B.2. An Example Implementation of SAML Based on the Cosine Metric

This section provides an example implementation of SAML based on the cosine metric as Figure 8.

B.3. Additional Illustrations of Semantically Relevant Local Regions

This section provides additional illustrations of semantically relevant local regions of a different category from the main text as Figure 9.

Figure 7. Some images from miniImageNet and CUB are shown for better understanding of datasets used in the experiments.

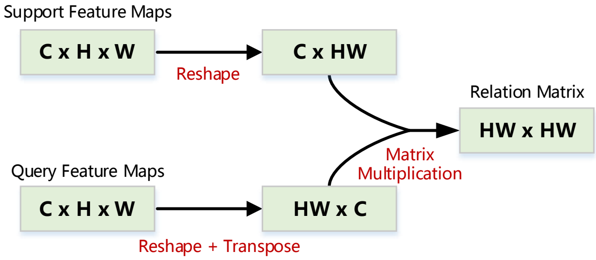

Figure 8. An example implementation of SAML based on the cosine metric.

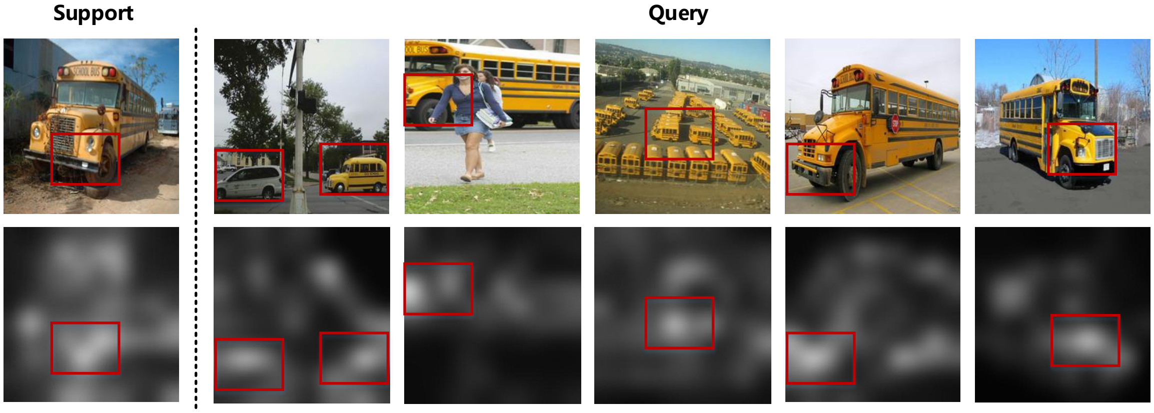

Figure 9. Illustrations of semantically relevant local regions. They demonstrate that the semantic alignment is realized by SAML.

B.4. Ablation Study of Metrics

Metrics can be induced by kernels. Specifically, kernel functions can first define inner products and then induce metrics. This section gives ablation studies on three more metrics other than the two included in the main text, cosine and Gaussian. They are respectively induced by the following three kernel functions.

Hyperbolic Tangent Kernel (HTK): It is defined as the following equation,

$$k(x,y) = \tanh(\alpha x^T y + c) = \frac{e^{\alpha x^T y + c} - e^{-\alpha x^T y - c}}{e^{\alpha x^T y + c} + e^{-\alpha x^T y - c}}. \tag{B.1}$$

Similarly, by replacing $\tanh$ with other nonlinear functions, such as ReLU and Sigmoid, we can get the following variants.

Non-Negative ReLU Kernel (NNRK): It is defined as the following equation,

$$k(x,y) = \mathrm{ReLU}(\alpha x^T y + c) = \begin{cases} \alpha x^T y + c, & \alpha x^T y + c > 0, \\ 0, & \alpha x^T y + c \le 0. \end{cases} \tag{B.2}$$

It is worth noting that this kernel is non-negative as its name suggests.

Non-Negative Sigmoid Kernel (NNSK): It is defined as the following equation,

$$k(x,y) = \text{Sigmoid}(\alpha x^T y + c) = \frac{1}{1+e^{-\alpha x^T y - c}}. \tag{B.3}$$

It is worth noting that this kernel is non-negative as its name suggests.

The experimental results of the ablation studies for these metrics are shown in Table 6 and Table 7 which are respective conducted on datasets miniImageNet and CUB. Similar to cosine and Gaussian, the attentional versions of these metrics perform much better than their non-attentional version. Also as shown in Table 6 and Table 7, our method SAML with the three new metrics all significantly outperform the state-of-the-art methods (for more details, please refer to Table 4 and Table 5 in the main text).

B.5. Comparisons with the State-of-the-art

All the exprimental results given in the main text is based on the embedding network adopted from a four-layer neural network. We also conducted experiemnts on a much deeper network, WRN-28 [45]. SAML with the metric of cosine distance still outperforms the state-of-the-art methods. Please see the results shown in Table 8.

Metric Functions	5way-1shot	5way-5shot
HTK	52.77±0.20%	68.83±0.16%
NNRK	53.93±0.20%	68.81±0.16%
NNSK	53.67±0.19%	69.16±0.47%
HTK + Attention	55.29±0.20%	**71.54±0.16%**
NNRK + Attention	55.84±0.20%	71.17±0.16%
NNSK + Attention	**56.10±0.20%**	71.51±0.16%

Table 6. The effect of different metric functions on the few-shot classification accuracies. Experiments are conducted on miniImageNet.

Metric Functions	5way-1shot	5way-5shot
HTK	63.61±0.22%	68.83±0.16%
NNRK	65.78±0.22%	79.62±0.16%
NNSK	64.12±0.23%	78.80±0.17%
HTK + Attention	70.05±0.22%	**81.97±0.15%**
NNRK + Attention	**70.13±0.21%**	81.38±0.16%
NNSK + Attention	69.67±0.21%	81.54±0.15%

Table 7. The effect of different metric functions on the few-shot classification accuracies. Experiments are conducted on CUB.

Model	5way-1shot	5way-5shot
Meta-SGD [22]	54.24±0.03%	70.86±0.04%
SNAIL [26]	55.71±0.99%	68.88±0.92%
Gidaris & Komodakis [10]	56.20±0.86%	73.00±0.64%
Matthias et al. [3]	56.30±0.40%	73.90±0.30%
Munkhdalai et al. [28]	57.10±0.70%	70.04±0.63%
TADAM [30]	58.50±0.30%	76.70±0.30%
Qiao et al. [33]	59.60±0.41%	73.74±0.19%
LEO [36]	61.76±0.08%	77.59±0.12%
SAML/cosine (ours)	**61.86±0.20%**	**77.68±0.15%**

Table 8. Few-shot classification accuracies on miniImageNet. Note that deeper networks are used here.