Supplementary Materials: Accurate and Lightweight Learning for Specific Domain Image-Text Retrieval

Anonymous Authors

1 REVIEW OF CLIP

CLIP consists of a dual-stream model comprising a text encoder and an image encoder. The feature extraction process is demonstrated by the following equations:

$$R_t = E_t(t) \tag{1}$$

$$R_v = E_v(v) \tag{2}$$

E_t and E_v represent the text encoder and image encoder, respectively. CLIP utilizes the output of the $< eos >$ token from the text encoder to represent the text feature R_t, and the output of the $< image >$ token from the image encoder to represent the image feature R_i. t and v represent the input text and image, respectively.

Both the text encoder (E_t) and image encoder (E_v) are constructed using stacked Transformer blocks. The Transformer block is defined as follows:

$$A_l = LN(ATT(H_l) + H_l) \tag{3}$$

$$H_{l+1} = LN(FFD(A_l) + A_l) \tag{4}$$

In the above equations, H_l represents the input sequence of the l-th Transformer block, ATT denotes the multi-head attention operation, LN stands for the layer normalization operation, and FFD represents the feedforward propagation using a fully connected layer.

For $l = 1$, H_1 is the input sequence consisting of the embedding of word (or image patch) concatenated with the $< eos >$ token (or $< image >$ token), along with positional encoding. CLIP uses a simple cross-modal contrastive loss L^{vtc} to optimize the learning of the model.

$$\begin{aligned} L^{vtc} = -\frac{1}{2}[\frac{1}{m}\sum_{j=1}^{m}\log\frac{exp\left(D\left(\mathbf{R_v}^j,\mathbf{R_t}^j\right)/\tau\right)}{\sum_{u=1}^{m}exp\left(D\left(\mathbf{R_v}^j,\mathbf{R_t}^u\right)/\tau\right)}+ \\ \frac{1}{m}\sum_{j=1}^{m}\log\frac{exp\left(D\left(\mathbf{R_t}^j,\mathbf{R_v}^j\right)/\tau\right)}{\sum_{u=1}^{m}exp\left(D\left(\mathbf{R_t}^j,\mathbf{R_v}^u\right)/\tau\right)}] \end{aligned} \tag{5}$$

Where L^{vtc} denotes the cross-modal contrastive loss, m denotes the batch size, D denotes the dot product similarity of vectors, $\mathbf{R_v} = \{R_v^i, i = 1, 2, \ldots\}$ is a matrix of image feature vectors within a batch, $\mathbf{R_t} = \{R_t^i, i = 1, 2, \ldots\}$ denotes a matrix of text feature vectors within a batch, and τ denotes the temperature coefficient.

2 PARAMETERS SEARCH OF MLCE LOSS

To further analyze the impact of the MLCE loss, we performed an additional parameter search for μ and α in Equations 3 and 4 on the RSITMD dataset, aiming to find the optimal parameter settings. We use the mR as the metric (Mean_R in the Figure). During the search for μ, we empirically set α to 1. The search was performed over the values {0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2}, and the results are shown in Figure 1. The optimal mR value of 50.22 was achieved when μ was set to 1 in ALR(w/o SPDS). Based on the μ search, we set μ to 1 and proceeded with the search for α. The search encompassed the values {0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000}, and the results are depicted in Figure 2. When α was set to 1,

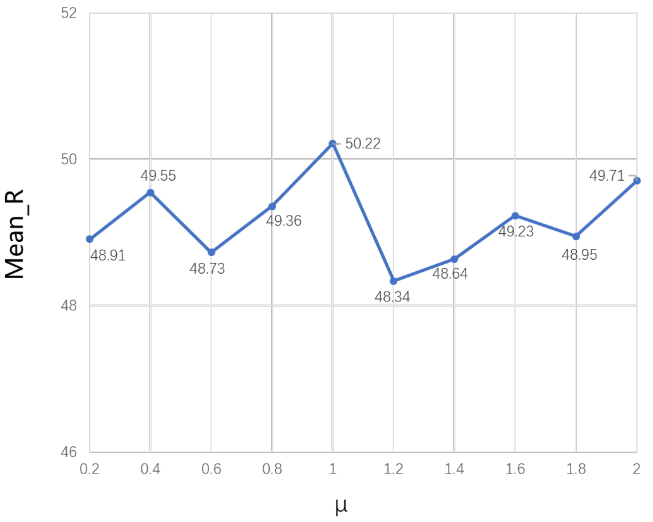

Figure 1: Parameter search results on the RSITMD dataset for μ.

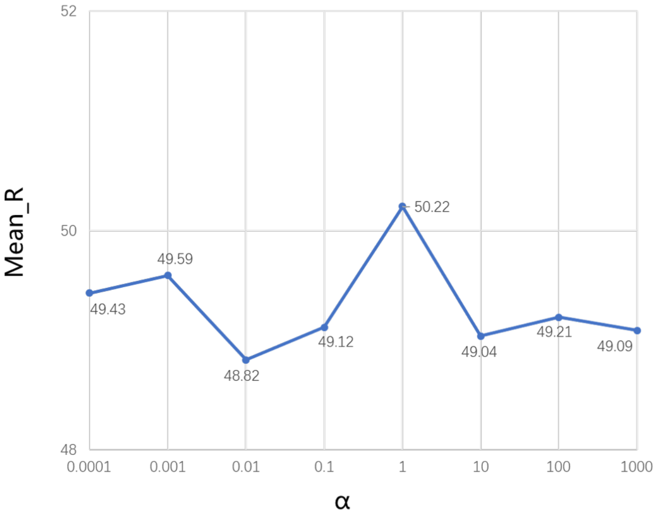

Figure 2: Parameter search results on the RSITMD dataset for α.

Table 1: Experimental results on UCM Caption

Method	Sentence Retrieval			Image Retrieval			*mR*	*Test FLOPs(G)*	*Test Pramaterrs(M)*
	R@1	*R@5*	*R@10*	*R@1*	*R@5*	*R@10*			
Traditional methods									
VSE++ [4] BMVC'18	12.38	44.76	65.71	10.10	31.80	56.85	36.93	2.44	15.78
SCAN [8] ECCV'18	14.29	45.71	67.62	12.76	50.38	77.24	44.67	2.42	13.68
CAMP [14] ICCV'19	14.76	46.19	67.62	11.71	47.24	76.00	43.92	2.28	36.64
MTFN [13] MM'19	10.47	47.62	64.29	14.19	52.38	78.95	44.65	2.80	77.90
RSITR methods									
AMFMN [20] TGRS'22	16.67	45.71	68.57	12.86	53.24	79.43	46.08	2.75	35.94
LW-MCR [21] TGRS'21	18.10	47.14	63.81	13.14	50.38	79.52	45.35	0.46	1.65
CABIR [23]	15.17	45.71	72.85	12.67	54.19	89.23	48.30	–	–
SSJDN [22] ACM TOMM'23	17.86	53.57	72.02	20.54	62.56	82.98	51.59	–	–
SMLGN [3] TGRS'24	12.86	49.52	75.71	14.29	52.76	84.67	48.30	–	–
MSITA [2] TGRS'24	16.86	49.33	73.33	14.29	57.16	91.58	50.43	–	–
Additional variants									
VIT + BERT	11.90	52.38	72.38	14.29	56.00	85.81	48.79	19.50	171.29
ResNet18 + BERT	12.86	43.81	68.57	10.29	50.86	81.62	44.67	4.46	97.28
ResNet50 + BERT	11.90	50.00	71.43	11.62	52.76	89.14	47.81	6.77	110.14
ResNet101+ BERT	16.67	53.33	76.19	15.52	51.14	76.38	48.21	10.50	129.13
CLIP based methods									
CLIP (zero-shot) [10]	8.57	34.76	62.86	9.43	39.43	66.95	37.00	13.21	82.46
CLIP (full-finetune) [10]	20.95	59.05	83.81	19.14	65.33	94.95	57.20	13.21	82.46
Maple [7]	18.57	55.24	80.00	16.67	63.04	94.00	54.60	13.21	86.79
OURS									
ALR (w/o SPDS)	22.38	60.48	81.90	19.71	67.14	94.76	**57.73**	13.21	82.46
ALR(k=9)	17.14	54.29	78.10	17.81	63.52	97.52	54.73	9.94	61.99
ALR(k=5)	14.29	47.62	69.05	13.81	52.67	82.67	46.68	5.57	34.70

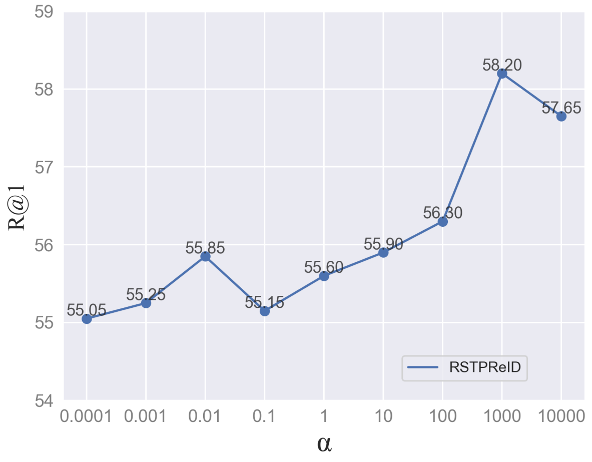

Figure 3: Parameter search results on the RSTPReID dataset for α.

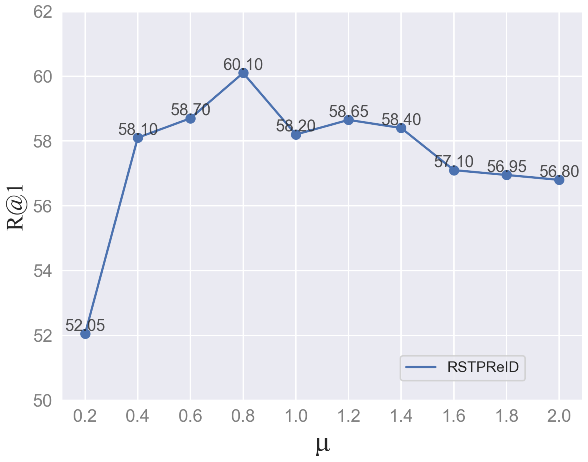

Figure 4: Parameter search results on the RSTPReID dataset for μ.

ALR(w/o SPDS) achieved the optimal mR value of 50.22. Therefore, in subsequent RSITMD experiments, we set both μ and α to 1.

In the TIReID task, we performed a parameter search for μ and α on the RSTPReID dataset, using R@1 as the evaluation metric. We set μ to 1 and proceeded with the search for α. This search included the values 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, and the results are depicted in Figure 3. When α was set to 1000, ALR(w/o SPDS) achieved the optimal R@1 value of 58.20. During the search for μ, we set α to 1000. The search encompassed the values {0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2}, and the results are shown in Figure 4. The optimal R@1 value of 60.10 was achieved when μ was set to 0.8 in ALR(w/o SPDS). Therefore, in subsequent TIReID experiments, we set μ to 0.8 and α to 1000.

Table 2: Experimental results on RSTPReID

Method	*R@1*	*R@5*	*R@10*	*mAP*	*Test FLOPs(G)*	*Test Parameters (M)*	*Ref*
Traditional methods							
C2A2 [9]	51.55	76.75	85.15	–	12.87	107.71	MM 22
LCR [19]	54.95	76.65	84.70	–	4.64	87.82	MM 23
Unipt [11]	49.45	72.75	80.35	–	13.59	142.91	ICCV 23
IVT [12]	46.70	70.00	78.80	–	19.07	142.32	ECCVW 22
TGDA [5]	48.35	73.15	80.30	37.96	5.08	113.84	TCSVT 23
IMG-NET [18]	37.60	61.15	73.55	–	–	–	JEI 20
AMEN [15]	38.45	62.40	73.80	–	–	–	PRCV 21
DSSL[24]	39.05	62.60	73.95	–	4.70	53.89	MM 21
SUM [16]	41.38	67.48	76.48	–	5.29	39.11	KBS 22
LBUL [17]	45.55	68.20	77.85	–	8.88	57.99	MM 22
CLIP based methods							
CLIP-full-finetune	55.05	79.25	86.55	43.68	13.21	84.46	ICML 21
CLIP-zero-shot	13.40	25.00	33.65	9.55	13.21	84.46	ICML 21
Ours							
ALR (w/o) SPDS	60.10	80.95	87.60	47.07	13.21	84.46	–
ALR (k=9)	45.35	68.05	78.60	37.03	9.94	61.99	–

3 EXPERIMENT RESULTS ON UCM CAPTION DATASET

Table 1 illustrates the results on the UCM Caption dataset. Our method continues to achieve the SOTA performance on this dataset. ALR (w/o SPDS) achieved an mR value of 57.73, while the lightweight versions, ALR (k=9) and ALR (k=5), achieved mR values of 54.73 and 46.68, respectively. Compared to traditional and RSITR methods, our ALR (w/o SPDS) outperformed the highest SSJDN method by 6.14. The lightweight ALR (k=9) also outperformed SSJDN by 3.14. Compared to additional variants, our method also demonstrated superiority. ALR (w/o SPDS) and ALR (k=9) achieved significant improvements in mR compared to these methods, while ALR (k=9) had much lower parameter count and computational cost than ResNet101 + BERT and VIT + BERT. Compared to CLIP-based methods, ALR (w/o SPDS) outperformed CLIP (full-finetune) by 0.53 in mR, highlighting the effectiveness of MLCE loss on UCM Caption. The lightweight ALR (k=9) achieved an acceptable trade-off between accuracy and computational cost (parameter count) compared to CLIP (full-finetune), maintaining high retrieval accuracy while significantly reducing computational cost and parameter count.

4 EXPERIMENT RESULTS ON RSTPREID DATASET

The comparative experimental results on the RSTPReID dataset are shown in Table 2. It is evident that ALR (w/o SPDS) achieved the highest R@K (K=1,5,10) scores and mAP values, surpassing all comparison algorithms. Compared to traditional methods, our ALR (w/o SPDS) outperformed the state-of-the-art LCR and C2A2 algorithms, while requiring fewer parameters. In comparison with CLIP-based methods, ALR (w/o SPDS) exceeded CLIP-full-finetune by 5.05 in R@1 and 3.39 in mAP, demonstrating the effectiveness of the MLCE loss. Our lightweight version, ALR(k=9), maintained competitive performance metrics while reducing computational and parameter requirements.

5 LAYER ABLATION OF SPDS ON REMAINING DATASETS

The results on three RSITR datasets are shown in Figure 5 - 7. From Figure 5, we can see that when K is set to 3, our self-distillation method achieved 33.65 on the RSITMD dataset, which surpassed the most of traditional comparison methods. We also found that as the number of layers increases, the mR achieved by the self-distillation method generally increases, which is intuitive. From Figure 6, we can see that mR reached 46.68 when K is set to 5, which also surpassed the traditional comparison methods on the UCM dataset. On the SYDNEY dataset, the advantage of the self-distillation method is not as significant as on other RSITR datasets, and it requires 9 layers to achieve a value of 52.76. On the UCM dataset, the SPDS achieved promising performance, while on the SYDNEY dataset, the improvement was not as significant. This is because the SYDNEY dataset is relatively small, with only 613 images, which can easily lead to overfitting. Therefore, the effect of network complexity on its performance is not very sensitive, meaning that the difference between large-scale and small-scale networks is not significant.

In the context of the TIReID task, K was set to {3, 4, 5, 6, 7, 8, 9, 10} and γ to {25, 30, 35, 40, 45}. Considering the task's emphasis on the precision of the first returned result, R@1 was selected as the evaluation metric. The results on the RSTPReid and CUHK PEDES datasets are depicted in Figure 8-9. The findings indicate that an K value of 8 maintains an R@1 at a high level, surpassing the comparative algorithms while significantly reducing computational and parameter requirements.

Building upon the aforementioned insights, it is evident that the SPDS significantly enhances efficiency, reducing parameter count and computational demands of large-scale models without compromising retrieval performance.

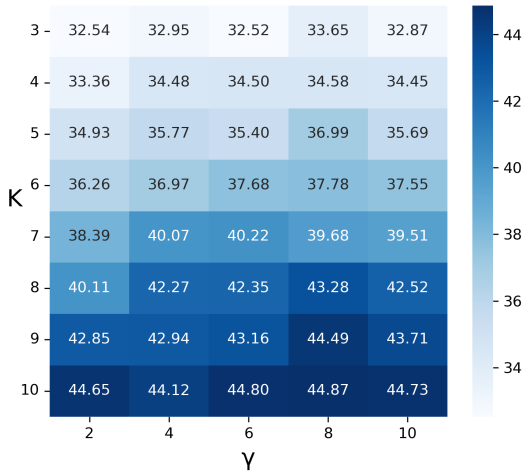

Figure 5: Joint search results for K and γ on RSITMD

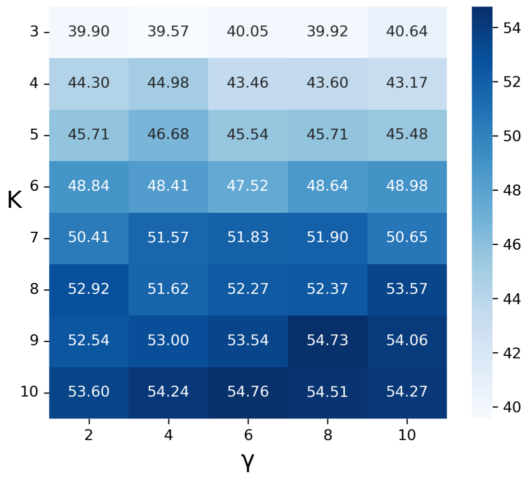

Figure 6: Joint search results for K and γ on UCM

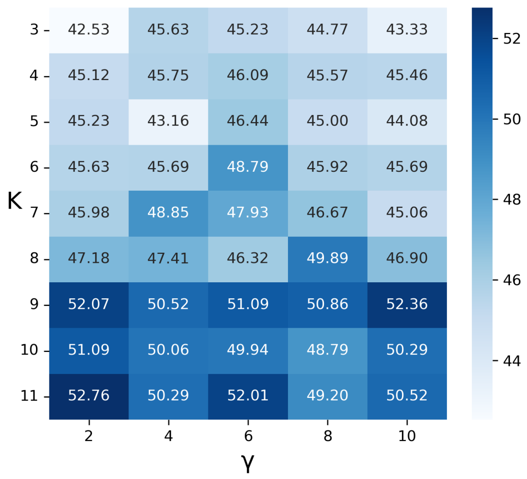

Figure 7: Joint search results for K and γ on Sydney

6 PARAMETERS CAPACITY AND COMPUTATIONAL COMPLEXITY ANALYSIS OF SPDS

We analyzed SPDS for the parameters capacity and computational complexity. We provide curves showing the variation of model parameters capacity and computational complexity with the number of layers, K. From Figure 10 and 11 , it can be observed that both the parameters capacity and computational complexity exhibit a linear growth trend as K increases. By employing the SPDS, parameters capacity and computational complexity can be significantly reduced. This reduction in computational complexity and parameters capacity is substantial compared to the CLIP-based methods, while still maintaining competitive retrieval performance.

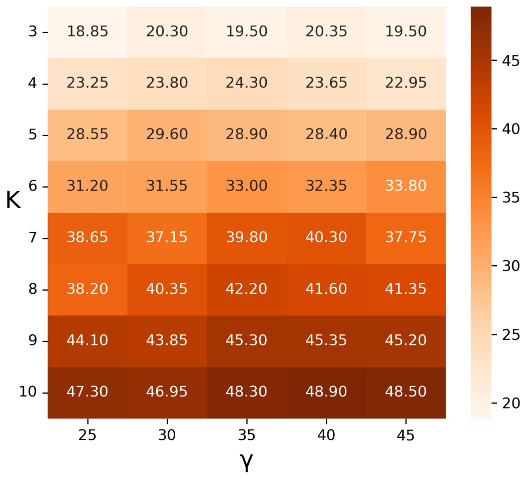

Figure 8: Joint search results for K and γ on RSTPReID

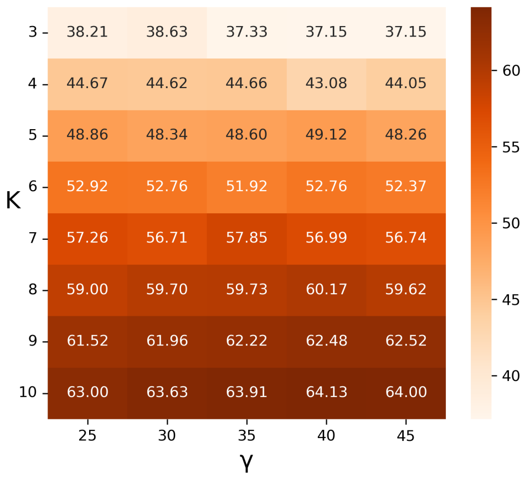

Figure 9: Joint search results for K and γ on CUHK-PEDES

7 STUDENT-TEACHER DISTILLATION

To conduct a comprehensive comparative analysis between the SPDS and traditional teacher-student distillation (STU-TEA), we designed a student model. By utilizing our ALR(w/o SPDS) as the teacher model, we implemented the conventional student-teacher knowledge distillation.

The student model consists of two branches: the image branch and the text branch. The image branch utilizes the Mobilenet-V3 model [6] as the feature extractor. The Mobilenet-V3 model comprises multiple bottlenecks, each incorporating channel separable convolution, channel attention structure, and residual connectivity. The model structure of Mobilenet-V3 is illustrated in Figure 12. Drawing inspiration from [1], the text branch employs multicore

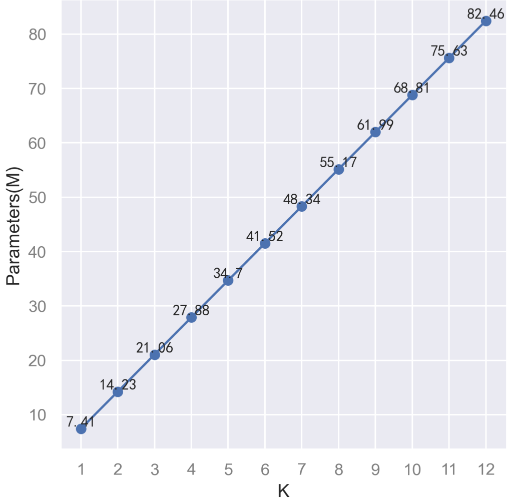

Figure 10: Curve of parameters capacity variation with K

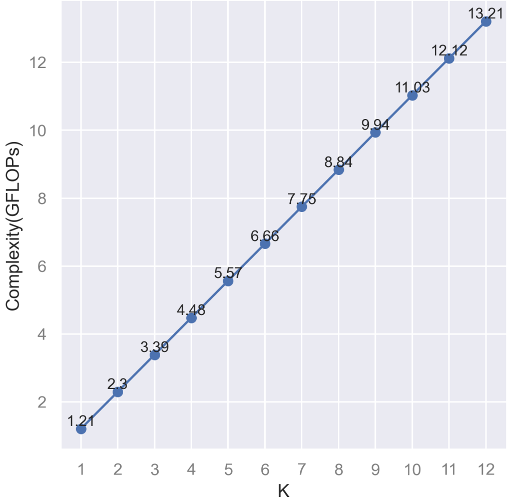

Figure 11: Curve of computation cost variation with K

convolution for feature extraction. Multicore convolution is a text processing model that offers parameter capacity and computation cost advantages over RNN-based and Transformer-based models. The lightweight image and text features with a batch produced by Mobilenet-V3 and multicore convolution are denoted as $\mathrm{Z_v}$ and $\mathrm{Z_t}$, respectively.

During student-teacher knowledge distillation, we first optimize the CLIP with L^{MLCE} (ALR w/o SPDS) to get the teacher model, and then fix the teacher model to optimize the lightweight student model. We design two losses when optimizing the student model, the contrastive loss L_{stu}^{vtc} of the student model and the cross-modal similarity knowledge distillation loss L_{stu}^{kd}.

$$
\begin{aligned}
L_{stu}^{vtc} = & -\frac{1}{2}[\frac{1}{m}\sum_{j=1}^{m}\log\frac{exp\left(D\left(\mathrm{Z_v}^{j},\mathrm{Z_t}^{j}\right)/\tau\right)}{\sum_{u=1}^{m}exp\left(D\left(\mathrm{Z_v}^{j},\mathrm{Z_t}^{u}\right)/\tau\right)}+ \\
& \frac{1}{m}\sum_{j=1}^{m}\log\frac{exp\left(D\left(\mathrm{Z_t}^{j},\mathrm{Z_v}^{j}\right)/\tau\right)}{\sum_{u=1}^{m}exp\left(D\left(\mathrm{Z_t}^{j},\mathrm{Z_v}^{u}\right)/\tau\right)}]
\end{aligned} \tag{6}
$$

m denotes the batch size, D denotes the dot product similarity of vectors, and λ denotes the temperature coefficient.

Then we calculate student-teacher knowledge distillation loss L_{stu}^{kd}. The cross-modal similarity matrix $\mathrm{S_3}$ needs to be calculated when calculating L_{stu}^{kd}: The formula is as follows:

$$
\mathrm{S_3} = \mathrm{Z_v}\mathrm{Z_t}^{T} \tag{7}
$$

T denotes the transpose operation.

$$
\begin{aligned}
L_{stu}^{kd} = (& -\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{exp\left(\mathrm{S_2}^{ij}/\theta\right)}{\sum_{k=1}^{m}exp\left(\mathrm{S_2}^{ik}/\theta\right)} \\
& log(\frac{exp\left(\mathrm{S_3}^{ij}/\theta\right)}{\sum_{k=1}^{m}exp\left(\mathrm{S_3}^{ik}/\theta\right)})) + \\
(& -\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{exp\left(\mathrm{S_2^T}^{ij}/\theta\right)}{\sum_{k=1}^{m}exp\left(\mathrm{S_2^T}^{ik}/\theta\right)} \\
& log(\frac{exp\left(\mathrm{S_3^T}^{ij}/\theta\right)}{\sum_{k=1}^{m}exp\left(\mathrm{S_3^T}^{ik}/\theta\right)}))
\end{aligned} \tag{8}
$$

$\mathrm{S_2}$ is the similarity matrix output by ALR w/o SPDS. θ is the distillation temperature coefficient.

The final loss of the student model is as follows.

$$
L^{student} = L_{stu}^{vtc} + \beta L_{stu}^{nk} \tag{9}
$$

where $L^{student}$ denotes the total loss of the student model and β denotes the combination coefficient.

8 PARAMETER SEARCH AND COMPARATIVE ANALYSIS OF STU-TEA

We also conducted a parameter search. We searched for the temperature coefficient θ of the distillation loss L_{stu}^{nk} and the combination coefficient β of the loss. We conducted parameter search experiments on RSITMD. We first conducted a search for the temperature coefficient θ, with the combination coefficient β fixed empirically at 1. We set θ={0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0}. The results are shown in Table 3. From Table 3, we can see that the optimal value of mR is achieved when θ is set to 0.4, and mR decreases when θ is too large or too small.

Next, we conducted a search for the combination coefficient β, with θ set to 0.4. We set β={0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000}. From Table 4, we can see that the optimal value of mR is achieved when β is set to 1000.

We examine the variation of mR values concerning K for SPDS on the RSITMD and compare them with ALR and STU-TEA. The experimental results are shown in Figure 13. We observed that as

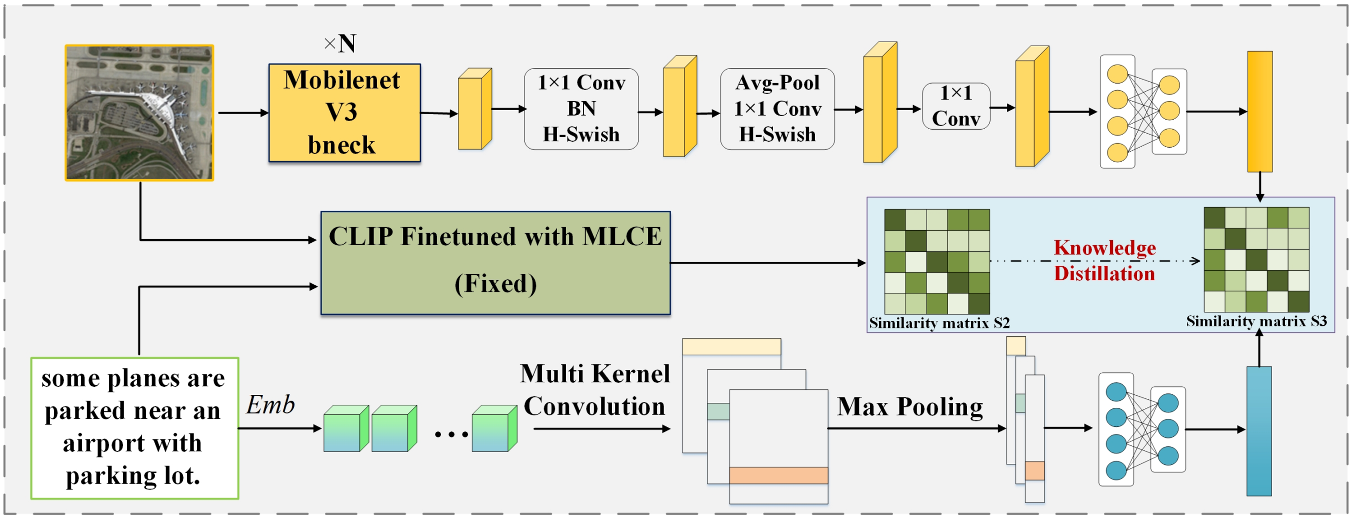

Figure 12: Student-teacher distillation

Table 3: Search results for distillation temperature coefficient θ in Student-teacher appproach on RSITMD

θ	0.2	0.4	0.6	0.8	1.00	1.2	1.4	1.6	1.8	2.00
mR	34.31	37.51	35.12	33.21	32.43	32.09	30.98	29.17	28.31	27.91

Table 4: Search results for combination coefficient β in Student-teacher appproach on RSITMD

β	0.0001	0.001	0.01	0.1	1	10	100	1000
mR	18.19	19.51	23.27	31.32	37.51	37.51	37.85	37.88

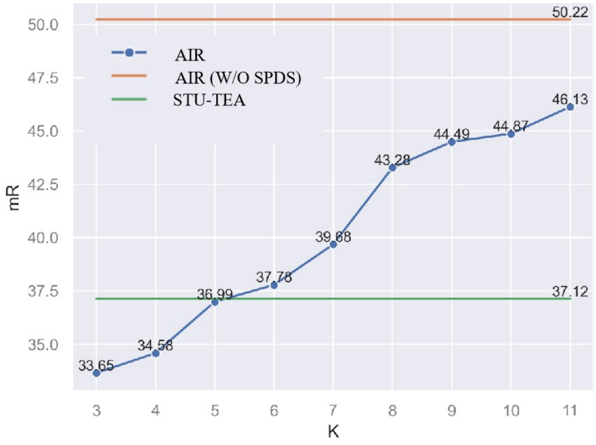

Figure 13: Curve of mR variation with K on RSITMD

K increases, the mR values for ALR also increase. On the RSITMD dataset, when K is greater than 5, ALR achieves higher mR values than STU-TEA. Additionally, ALR(W/O SPDS) with 12 Transformer blocks has the highest mR values on both datasets, reaching 50.22.

REFERENCES

[1] Yahui Chen. 2015. *Convolutional neural network for sentence classification.* Master's thesis. University of Waterloo.

[2] Yaxiong Chen, Jinghao Huang, Xiaoyu Li, Shengwu Xiong, and Xiaoqiang Lu. 2023. Multiscale Salient Alignment Learning for Remote Sensing Image-Text Retrieval. *IEEE Transactions on Geoscience and Remote Sensing* (2023).

[3] Yaxiong Chen, Jirui Huang, Shengwu Xiong, and Xiaoqiang Lu. 2024. Integrating Multisubspace Joint Learning With Multilevel Guidance for Cross-Modal Retrieval of Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–17.

[4] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).

[5] Liying Gao, Kai Niu, Bingliang Jiao, Peng Wang, and Yanning Zhang. 2023. Addressing information inequality for text-based person search via pedestrian-centric visual denoising and bias-aware alignments. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[6] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision.* 1314–1324.

[7] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. MaPLe: Multi-Modal Prompt Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 19113–19122.

[8] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV).* 201–216.

[9] Kai Niu, Linjiang Huang, Yan Huang, Peng Wang, Liang Wang, and Yanning Zhang. 2022. Cross-modal co-occurrence attributes alignments for person search by language. In *Proceedings of the 30th ACM International Conference on Multimedia.* 4426–4434.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning.* PMLR, 8748–8763.

[11] Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. 2023. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 11174–11184.

[12] Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. 2022. See finer, see more: Implicit modality alignment for text-based person retrieval. In *European Conference on Computer Vision.* Springer, 624–641.

[13] Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. 2019. Matching images and text with multi-modal tensor fusion and re-ranking. In *Proceedings of the 27th ACM international conference on multimedia*. 12–20.

[14] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[15] Zijie Wang, Jingyi Xue, Aichun Zhu, Yifeng Li, Mingyi Zhang, and Chongliang Zhong. 2021. Amen: Adversarial multi-space embedding network for text-based person re-identification. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part II 4*. Springer, 462–473.

[16] Zijie Wang, Aichun Zhu, Jingyi Xue, Daihong Jiang, Chao Liu, Yifeng Li, and Fangqiang Hu. 2022. SUM: Serialized Updating and Matching for text-based person retrieval. *Knowledge-Based Systems* 248 (2022), 108891.

[17] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. 2022. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1984–1992.

[18] Zijie Wang, Aichun Zhu, Zhe Zheng, Jing Jin, Zhouxin Xue, and Gang Hua. 2020. IMG-Net: inner-cross-modal attentional multigranular network for description-based person re-identification. *Journal of Electronic Imaging* 29, 4 (2020), 043028–043028.

[19] Shuanglin Yan, Neng Dong, Jun Liu, Liyan Zhang, and Jinhui Tang. 2023. Learning comprehensive representations with richer self for text-to-image person re-identification. In *Proceedings of the 31st ACM international conference on multimedia*. 6202–6211.

[20] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. 2021. Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–19.

[21] Zhiqiang Yuan, Wenkai Zhang, Xuee Rong, Xuan Li, Jialiang Chen, Hongqi Wang, Kun Fu, and Xian Sun. 2021. A lightweight multi-scale crossmodal text-image retrieval method in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–19.

[22] Chengyu Zheng, Ning Song, Ruoyu Zhang, Lei Huang, Zhiqiang Wei, and Jie Nie. 2023. Scale-semantic joint decoupling network for image-text retrieval in remote sensing. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 1 (2023), 1–20.

[23] Fuzhong Zheng, Weipeng Li, Xu Wang, Luyao Wang, Xiong Zhang, and Haisu Zhang. 2022. A Cross-Attention Mechanism Based on Regional-Level Semantic Features of Images for Cross-Modal Text-Image Retrieval in Remote Sensing. *Applied Sciences* 12, 23 (2022), 12221.

[24] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 209–217.