Sketch-Based Poetry Retrieval with Unsupervised Vision-and-Language Pre-training

Yuqing Li[1], Yuting Wei[1], Yangfu Zhu[2], and Bin Wu[1](✉)

[1] Beijing University of Posts and Telecommunications, Beijing, China
{liyuqing, yuting_wei, wubin}@bupt.edu.cn
[2] Capital Normal University, Beijing, China
b556@cnu.edu.cn

Abstract. *"Poetry in pictures and pictures in poetry"* elucidates a naturally direct correlation between the classical Chinese poetry and vision, making the image and poetry retrieval crucial. Most existing works employ the photograph as the query, but obtaining a suitable photograph brings additional difficulties. In contrast, the free-hand sketch has served as a more convenient tool to depict human perception since ancient times. In this paper, we introduce a new task of Sketch-Based Poetry Retrieval. The task is challenging due to the following factors: (i) the significant differences between sketches and images, as well as between poetry and modern Chinese; (ii) the high time cost of collecting the parallel cross-modal data for the traditional supervised learning. To address these challenges, we construct a sketch-and-poetry pre-training model based on unsupervised vision-and-language learning named **SKP-CLIP**. Specifically, we utilize a multi-modal knowledge graph for poetry to bridge the semantic gap and then learn the modal alignment through sketch and its caption as well as poetry and its corresponding image entities instead of sketch-poetry pairs. Furthermore, we semi-automatically assemble a dataset for evaluation. Experiments confirm the effectiveness of our method and establish benchmarks for this new task.

Keywords: Multi-modal Retrieval · Knowledge Enhance · Low Resource.

1 Introduction

Classical Chinese poetry, as an invaluable cultural heritage, has high artistic value and serves as an essential medium to record history and express emotions since ancient times. The poetry is distinct from the straightforward text commonly encountered, possessing a unique visual charm. A well-known adage *"Poetry in pictures and pictures in poetry"*, emphasizing the natural connection between the classical Chinese poetry and vision. Simultaneously, vision, as a cross-cultural mode of communication, contributes to the comprehension of classical Chinese poetry with complex grammar and semantics. Consequently, the image and poetry retrieval can enable individuals to acquire and understand poetry from a more flexible, suitable, and intuitive perspective which makes this task stand as a crucial component of the poetry research. Existing works [11] mainly center on the photograph to poetry retrieval. But when an idea flashes, there is still a certain threshold to obtain suitable photographs. Unlike photographs, the free-hand sketch serves as a more accessible visual medium. Just like the popular pictionary game, *Quick, Draw!*, even children can express their complex ideas through a few simple strokes. In light of these, we introduce a novel task named **Sketch-Based Poetry Retrieval**, which is to take the sketch as the query to retrieve the relevant poetry from the candidates. It aims not only to advance the study of classical Chinese poetry but also to allure a wider audience to the world of poetry, leveraging the amusing factor of the sketch itself.

In the field of image and text retrieval, one of the most popular methods is based on the vision-and-language pre-training model. However, these models commonly require a massive amount of parallel data to learn the alignment of image and text. This needs such a considerable time cost that it is hard to collect and scale up in the field of sketch and poetry. To explore this problem, an Unsupervised Vision-and-Language Pre-training (UVLP) method that indirectly learns the alignment between sketch and poetry is a better choice. Existing studies on UVLP [18, 1, 8] leverage images and their object labels as the pseudo parallel data to learn cross-modal alignment. Due to the significant differences in the syntax and the distribution of objects, existing methods fail to be applied in the field of sketch and poetry.

In this paper, we propose **SKP-CLIP**, a **SK**etch-and-**P**oetry Pre-training model based on **CLIP** with a UVLP method enhanced by a multi-modal knowledge graph. In particular, SKP-CLIP engages in a two-stage alternating learning. In the sketch training stage, pairs from a sketch-image-caption dataset, FS-COCO [3], are used to guide the sketch representation closer to the context. In the poetry training stage, a multi-modal knowledge graph for poetry is employed to offer natural connections between text entities and image entities. We also introduce edge maps to serve as an intermediate form between the image and sketch. For the evaluation of this task, we semi-automatically construct a sketch-poetry dataset. Experiments on this dataset show that SKP-CLIP with such small-scale and non-parallel datasets performs better than those large-scale pre-training models, which proves the effectiveness of our method. We hope our method will promote the development of other low-resource multi-modal tasks.

In summary, the contributions of our work are as follows:

- We propose a novel task named Sketch-Based Poetry Retrieval, displaying classical Chinese poetry in a cross-cultural visual mode.
- We construct a sketch-and-poetry model with unsupervised vision-and-language pre-training (SKP-CLIP)[3], making it possible to bridge the gap between the sketch and poetry without parallel data.
- We release a test dataset for the sketch-based poetry retrieval task and provide various benchmarks. Experiments prove the effectiveness of our method.

[3] The code and dataset are in https://github.com/liyuqing1/SKP-CLIP.

2 Related Work

Vision-and-Language Pre-training (VLP). VLP models aim to map images and texts into a unified semantic vector space. Recently, the Chinese domain has witnessed the emergence of VLP models [4, 7, 15, 14, 16]. They can be categorized into two architectures: single-stream and dual-stream. Single-stream models fuse image and text features early for learning more cross-modal interaction. Dual-stream models independently encode images and texts so candidate features can be pre-calculated offline, powerfully accelerating inference. Thus, this paper focuses on dual-stream architecture due to its higher retrieval speed.

Unsupervised Vision-and-Language Pre-training (UVLP). The traditional VLP model demands millions of image-text pairs for supervised learning which is extremely time-consuming. Meanwhile, semi-supervised methods [2, 6] use the similarity between raw data and parallel anchor pairs as guidance but are difficult to apply due to the subjectivity of this task and the diversity of the sketch. To overcome this, the Unsupervised Vision-and-Language Pre-training (UVLP) model is proposed which pre-trains without any parallel data. Existing studies [18, 1, 8] utilize pseudo-parallel pairs of images and their object labels. However, the significant semantic gap between labels and poetry brings great challenges to sketch-and-poetry pre-training. Specifically, popular objects in ancient and modern times are inconsistent, e.g., the airplane is modern.

3 Methods

3.1 Model Overview

The overall architecture of SKP-CLIP is shown in Fig. 1. It takes CN-CLIP [16], a CLIP model pre-trained on Chinese datasets, as the backbone and is also initialized by CN-CLIP. The pre-training process is divided into two alternating stages. We also employ a multi-modal knowledge graph for poetry, PKG [11], to align poetry with vision and enhance the understanding of poetry.

3.2 Sketch Training Stage

In the sketch training stage, a sketch-image-caption dataset, FS-COCO [3], is adopted to align the sketch with Chinese text. Firstly, the caption is translated into modern Chinese through a translation tool, i.e., Baidu Translation[4]. Following [11], the related text entities are then inserted into the caption e.g., "aircraft" behind "airplane" in Fig. 1. The final text is encoded by the frozen poetry encoder to obtain the caption representation v_{caption}. The sketch is encoded into the sketch representation v_{sketch}, by the sketch encoder. The image is encoded into the image representation v_{image}, by the frozen teacher model CN-CLIP.

To learn the cross-modal representation of the sketch, we use two tasks: (i) IT (Image-Text contrast) task for sketch aims to align the representations of

[4] https://fanyi.baidu.com/

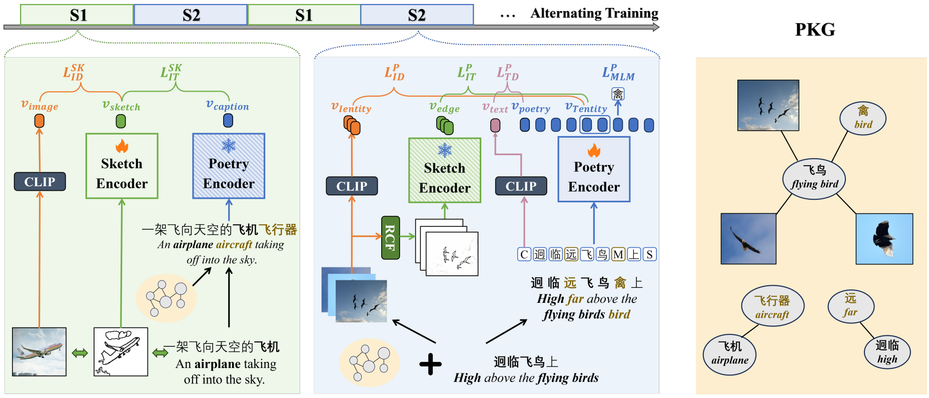


Fig. 1: The overview architecture of SKP-CLIP and the related part of KG.

sketch and caption. Due to the relatively small scale of sketches and the limited semantic discrimination of captions, we take the cosine loss as follows:

$$\mathcal{L}_{\text{IT}}^{\text{SK}} = 1 - \cos\left(v_{\text{sketch}}, v_{\text{caption}}\right). \tag{1}$$

(ii) ID (Image feature guide Distillation) task for sketch aims to constrain the sketch representation not to deviate from the rich visual semantic space of CN-CLIP. We use infoNCE loss to realize this task:

$$\mathcal{L}_{\text{ID}}^{\text{SK}} = -\frac{1}{2}\left(\log\frac{\exp\left(v_{\text{image}} \cdot v_{\text{sketch}_+}/\tau\right)}{\sum_{i=0}^{K}\exp\left(v_{\text{image}} \cdot v_{\text{sketch}}^{i}/\tau\right)} + \log\frac{\exp\left(v_{\text{sketch}} \cdot v_{\text{image}_+}/\tau\right)}{\sum_{i=0}^{K}\exp\left(v_{\text{sketch}} \cdot v_{\text{image}}^{i}/\tau\right)}\right). \tag{2}$$

where τ is a trainable temperature parameter, v_+ denotes the positive sample and K denotes the number of samples. The final optimization function of the sketch training stage $\mathcal{L}^{\text{SK}}$ is the weighted sum of these two losses defined as:

$$\mathcal{L}^{\text{SK}} = \alpha_1 \mathcal{L}_{\text{IT}}^{\text{SK}} + \alpha_2 \mathcal{L}_{\text{ID}}^{\text{SK}}. \tag{3}$$

3.3 Poetry Training Stage

In the poetry training stage, a text-only poetry corpus is used to guide the understanding of poetry. The difference from the sketch stage is that there's no visual data matched with the poetry. Thus, image entities connected with the inserted text entities from PKG are used as matched images instead. Firstly, text entities are inserted into the poetry like the sketch training stage. Then, image entities related to these text entities are collected and transformed into edge maps via RCF [13], which are intermediate forms between sketches and images. Assume the representations of the poetry and image entity encoded by CN-CLIP are v_{text} and v_{Ientity}, the edge map encoded by the frozen sketch encoder is v_{edge}, and the inserted text entity and [CLS] token encoded by the poetry encoder are v_{Tentity} and v_{poetry} respectively.

To learn the poetry representation fitting with the cross-modal semantic space, there are four tasks: (i) MLM (Mask Language Model) task learns the contextual semantics following the mask strategy of Whole Word Masking. (ii) IT (Image-Text contrast) task for poetry aims to learn the alignment between the poetry and edge map. Since a text entity commonly connects to multiple edge maps and consists of multiple tokens, we count the average of all involved token representations to obtain the representation of text entity v_{Tentity}, and calculate the cosine loss with min-pooling:

$$\mathcal{L}_{\text{IT}}^{\text{P}} = \min_{j} \left\{ 1 - \cos\left(v_{\text{edge}}^{j}, v_{\text{Tentity}}\right) \right\}. \tag{4}$$

(iii) ID (Image feature guide Distillation) and TD (Text feature guide Distillation) tasks are both devoted to preserving the image-text alignment from CN-CLIP. Due to the many-to-one distribution of image entities and poetry, the ID task also uses min-pooling of cosine loss:

$$\mathcal{L}_{\text{ID}}^{\text{P}} = \min_{j} \left\{ 1 - \cos\left(v_{\text{Ientity}}^{j}, v_{\text{Tentity}}\right) \right\}. \tag{5}$$

While the TD task uses infoNCE loss between v_{text} and v_{poetry} denoted as $\mathcal{L}_{\text{TD}}^{\text{P}}$:

$$\mathcal{L}_{\text{TD}}^{\text{P}} = -\frac{1}{2} \left(\log \frac{\exp\left(v_{\text{text}} \cdot v_{\text{poetry}_{+}}/\tau\right)}{\sum_{i=0}^{K} \exp\left(v_{\text{text}} \cdot v_{\text{poetry}}^{i}/\tau\right)} + \log \frac{\exp\left(v_{\text{poetry}} \cdot v_{\text{text}_{+}}/\tau\right)}{\sum_{i=0}^{K} \exp\left(v_{\text{poetry}} \cdot v_{\text{text}}^{i}/\tau\right)} \right). \tag{6}$$

The final loss of the poetry training stage $\mathcal{L}^{\text{P}}$ is the weighted sum of these losses:

$$\mathcal{L}^{\text{P}} = \beta_1 \mathcal{L}_{\text{MLM}}^{\text{P}} + \beta_2 \mathcal{L}_{\text{IT}}^{\text{P}} + \beta_3 \mathcal{L}_{\text{ID}}^{\text{P}} + \beta_4 \mathcal{L}_{\text{TD}}^{\text{P}}. \tag{7}$$

4 Experiments

4.1 Datasets

Training datasets. During pre-training, we need a sketch-image-caption dataset and a text-only poetry dataset. For the sketch-image-caption dataset, we select a scene-level dataset as opposed to an object-level one, due to the adaptability of the downstream application and the richer contextual information that scene-level sketches offer. There are some available scene-level sketch datasets, i.e., FS-COCO [3] and SketchyCOCO [5]. However, sketches in SketchyCOCO are automatically generated and only depict 17 objects. Thus, we choose FS-COCO which consists of 10,000 high-quality free-hand sketches each with an image and a caption, and contains 157 objects. The train, valid, and test set ratio is 7:1.5:1.5. As for the poetry dataset, we collect 257,568 poetry in diverse genres ranging from the pre-Qin period (around 1000 BC) to the Qing dynasty (up to 1912 AD). It is divided into train, valid, and test sets with a ratio of 8:1:1.

Sketch-Poetry Test dataset. We semi-automatically construct a sketch-poetry test dataset for the novel sketch-based poetry retrieval task as follows:

(1) Preprocessing: We take the test set of FS-COCO as the sketch query, and use the poetry-modern Chinese dataset, CCPM [10], as the poetry candidate. After standardization, there are 23,997 one-to-one pairs of poetry. (2) Indirect Match. Large-scale Chinese VLP models, i.e., Wukong [7], R2D2 [15], ERNIE-ViL2 [14] and CN-CLIP, are used to match images corresponding to sketches in FS-COCO with the modern translations of poetry in CCPM. A poetry-and-image pre-training model, PKG-BERT [11], is also applied to match images with the poetry. (3) Filter Candidates. We take the top 10 answers for each sketch and discard low-quality answers using the third quartile (Q3) as the threshold. Frequent occurrences of poetry lacking visual information introduce noise, so five experts are invited to verify those high-frequency answers. (4) Secondary Validation. We translate all object labels into Chinese and collect their synonyms. Answers without any label from the query sketch will be discarded. Through the process above, a sketch-poetry test dataset with 1,490 queries and an average of 6.17 objects per query is finally obtained.

4.2 Settings

Baselines. For the sketch-based poetry retrieval task, we compare SKP-CLIP with recent dual-stream Chinese VLP models, i.e., BriVL [4], Wukong, CN-CLIP, R2D2, ERNIE-ViL2, and a poetry-and-image pre-training model PKG-BERT. For fairness, we also try to translate the poetry into modern Chinese via Baidu Translation, to fit these baselines pre-trained on modern Chinese (denoted as “w/modern”). We adopt the metrics: HR@K (Hit Ratio) and F1@K (F1-score).

Implementation Details. The max length of the input poetry is 128. The learning rate of the sketch training stage is set to 5e-7 and the learning rate of the poetry training stage is set to 5e-5 after grid searches. The other settings are consistent with CN-CLIP. We adopt DWA [12] to dynamically adjust the weights of these losses i.e., α_1, α_2, β_1, β_2, β_3 and β_4, with a temperature of 2.

5 Results and Analysis

5.1 Overall Results

Table 1 reports results on the sketch-based poetry retrieval task. We can observe that: (1) SKP-CLIP outperforms all the baselines, particularly in comparison to the backbone model, CN-CLIP, which indicates that SKP-CLIP benefits from the proposed unsupervised vision-and-language pre-training method with the multi-modal knowledge graph. (2) The translation corpus brings slight or no improvement. This is because the translation of poetry is still a challenging task. (3) Except SKP-CLIP, the baseline pre-trained on a larger scale of parallel data seems to have better performance (i.e., ERNIE-ViL2 $>$CN-CLIP and R2D2 $>$Wukong $>$BriVL). It demonstrates the effectiveness of our UVLP method which enables the model to achieve enhanced performance in a novel domain, even pre-training with a limited-scale and non-parallel dataset.

Table 1: Overall results of the sketch-based poetry retrieval task where bold texts indicate the best results and underlined texts note the second-best results.

Model		HR@5	HR@10	HR@20	F1@5	F1@10	F1@20
BriVL	w/poetry	0.02953	0.04295	0.04295	0.00671	0.00552	0.00319
	w/modern	0.02550	0.04362	0.07651	0.00534	0.00516	0.00589
Wukong	w/poetry	0.05570	0.09262	0.14295	0.01396	0.01356	0.01410
	w/modern	0.04966	0.09530	0.14966	0.01150	0.01327	0.01343
CN-CLIP	w/poetry	0.12483	0.18658	0.27718	0.03449	0.03192	0.03267
	w/modern	0.12215	0.18255	0.27383	0.03244	0.02862	0.02848
R2D2	w/poetry	0.11409	0.17852	0.25436	0.02970	0.02848	0.03089
	w/modern	0.09799	0.17114	0.23826	0.02327	0.02582	0.02662
ERNIE-ViL2	w/poetry	0.14497	0.20604	0.27852	0.03777	0.03371	0.03159
	w/modern	0.12953	0.18993	0.26644	0.03449	0.02977	0.02898
PKG-BERT	w/poetry	0.05839	0.08792	0.08792	0.01588	0.01363	0.00788
SKP-CLIP	w/poetry	**0.21074**	**0.30671**	**0.38725**	**0.04913**	**0.04325**	**0.03860**

5.2 Ablation Study

Do Pre-training Tasks Work? To verify the effectiveness of each learning task, we construct the following variants: “w/o edge” removes the edge maps which supports the IT task for poetry $\mathcal{L}_{\mathrm{IT}}^{\mathrm{P}}$; “w/o MLM” removes the MLM task; “w/o distill” removes the distillation from CN-CLIP; “w/o KG” removes the insertion of PKG; “w/o sketch” only trains the model in the poetry stage; “w/o poetry” only trains the model in the sketch stage; and “Text-Only” simply fine-tunes the poetry encoder on the poetry corpus with $\mathcal{L}_{\mathrm{MLM}}^{\mathrm{P}}$ and $\mathcal{L}_{\mathrm{TD}}^{\mathrm{P}}$.

The results are reported in Table 2. Observations reveal the following: (1) Each variant has performance degradation, proving the necessity of each training task. (2) “w/o distill” performs worst, indicating that in cases of limited data scale, the distillation remains essential, which constrains the model from forgetting the prior knowledge. (3) Both “w/o sketch” and “w/o poetry” underperform, proving the necessity of the two-stage alternating training, and “w/o sketch” performs worse suggesting that vision is a greater challenge for this task. (4) “Text Olny” is even worse than “CN-CLIP”, indicating that fine-tuning only on the text side may unfortunately result in the forgetting of cross-modal alignment.

How to Pre-train the Model? Following the previous experience of freezing most of the image encoder [7], we train the last layer of the sketch encoder and the entire poetry encoder. To obtain an effective pre-training approach, we attempt the following variants: “V=T=1” only trains the last layer of both poetry and sketch encoder; “MLP2” adds two-layer MLP at the end of CLIP and only trains this MLP [17]; “MLP4” adds four-layer trainable MLP; “SPA”, inspired by the PEFT of LLMs [9], adds trainable adapters to each layer in the form of LoRA.

Results shown in Table 3 reveal: (1) Our method gains the best result proving its suitability for the current situation. (2) “MLP4” is better than “MLP2” indicating that the transfer of modalities may require more complex structures.

Table 2: Ablation study on pre-training tasks.

Model	HR@5	HR@10	HR@20	F1@5	F1@10
SKP-CLIP	**0.2107**	**0.3067**	**0.3872**	**0.0491**	**0.0433**
w/o edge	0.0993	0.1584	0.2396	0.0229	0.0218
w/o MLM	0.1208	0.1725	0.2523	0.0318	0.0279
w/o distill	0.0228	0.0396	0.0658	0.0048	0.0048
w/o KG	0.0651	0.1383	0.2161	0.0152	0.0174
w/o sketch	0.0477	0.0725	0.1107	0.0101	0.0085
w/o poetry	0.1436	0.2034	0.3034	0.0416	0.0367
Text-Only	0.1114	0.1705	0.2443	0.0270	0.0237

(3) "V=T=1" is worse than "V=1, T=all" which indicates that providing more trainable parameters on the text side aids in multi-modal learning. (4) "SPA" performs poorly, suggesting that changes inside the Transformer blocks may introduce more noise and are inapplicable for the vision-and-language pre-training.

Table 3: Ablation study on tuning strategies.

Model	HR@5	HR@10	HR@20	F1@5	F1@10
V=1,T=all	**0.2107**	**0.3067**	**0.3872**	**0.0491**	**0.0433**
V=T=1	0.1275	0.2013	0.3114	0.0304	0.0275
MLP2	0.1054	0.1658	0.2389	0.0248	0.0235
MLP4	0.1362	0.2154	0.2966	0.0342	0.0331
SPA	0.0953	0.1530	0.2221	0.0203	0.0182

What is the Effective Input? We pre-train SKP-CLIP on the following sketch-text pairs: "w/ sent" uses a caption without any text entities; "w/ label" only uses object labels; "w/ merge" merges the caption and object labels through prompt; "w/ FS+SK" uses sketches from both FS-COCO and SketchyCOCO.

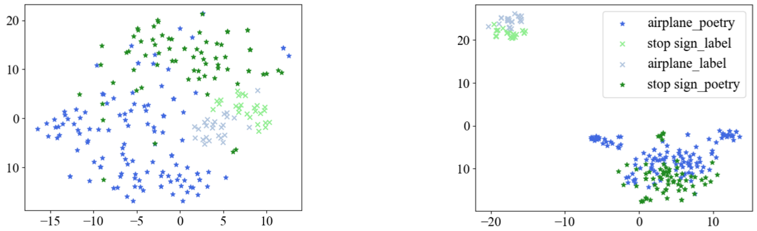
As shown in Table 4, we can observe: (1) SKP-CLIP (i.e., "w/ sent+KG") is superior to "w/ sent" demonstrating the necessity of incorporating KG into sketch captions. (2) "w/ label" is worse than "w/ sent", which indicates a significant disparity between the sequence of object labels and the complete sentence.(3) "w/ merge" gains the second-best results indicating that emphasizing keywords in the context can enhance the performance. In a sense, inserting text entities is a form of emphasis on keywords. (4) Despite the larger scale, "w/ FS+SK" performs poorly suggesting that it may value the quality of data more than the quantity, and a semi-automatic sketch dataset could introduce noise.

5.3 Visualization and Explanation

To intuitively comprehend the enhancement of our method, we visualize the distribution of features from SKP-CLIP and CN-CLIP via t-SNE on our test

Table 4: Ablation study on the input format.

Model	HR@5	HR@10	HR@20	F1@5	F1@10
w/ sent+KG	**0.2107**	**0.3067**	**0.3872**	**0.0491**	**0.0433**
w/ sent	0.0919	0.1336	0.2128	0.0233	0.0189
w/ label	0.0738	0.1215	0.1893	0.0164	0.0153
w/ merge	0.1450	0.2228	0.3114	0.0341	0.0298
w/ FS+SK	0.0255	0.0530	0.0906	0.0053	0.0063



(a) SKP-CLIP (our)

(b) CN-CLIP

Fig. 2: T-SNE visualization of SKP-CLIP and CN-CLIP.

dataset. Take two modern objects, “airplane” and “stop sign”, for instance. We obtain label features from object labels and their synonyms and encode poetry features from poetry containing relevant ancient words. As Fig. 2b, features of poetry and modern labels in CN-CLIP are separated, indicating its limited capacity for cross-temporal semantic mapping. As Fig. 2a, poetry and label features in SKP-CLIP are closer and also retain distinctiveness across different classes proving the connection between modern and ancient concepts is established.

6 Conclusion

In this paper, we propose a novel task, Sketch-Based Poetry Retrieval, binding classical Chinese poetry with a convenient and cross-cultural visual modality, and assemble a test dataset with benchmarks for this task. We construct a sketch-and-poetry model, SKP-CLIP, to bridge the gap between the sketch and poetry pre-training through an effective UVLP method based on a multi-modal knowledge graph without any sketch-poetry pair. Extensive experiments prove the effectiveness of our method. We also expect our method could encourage more exploration of the low-resource multi-modal studies.

Acknowledgments. This work is supported by the National Natural Science Foundations of China under Grant (62372060).

References

1. Chen, C., Li, P., Sun, M., Liu, Y.: End-to-end unsupervised vision-and-language pre-training with referring expression matching. In: Proceedings of the 2022 Con-

ference on Empirical Methods in Natural Language Processing. pp. 10799–10810 (2022)

2. Chen, C., Li, P., Sun, M., Liu, Y.: Weakly supervised vision-and-language pretraining with relative representations. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. pp. 8341–8355 (2023)
3. Chowdhury, P.N., Sain, A., Bhunia, A.K., et al.: FS-COCO: Towards understanding of freehand sketches of common objects in context. In: Computer Vision – ECCV 2022. pp. 253–270 (2022)
4. Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Lu, H., Song, R., Gao, X., Xiang, T., Sun, H., Wen, J.R.: Towards artificial general intelligence via a multimodal foundation model. Nature Communications **13**(1), 3094 (2022)
5. Gao, C., Liu, Q., Xu, Q., Wang, L., Liu, J., Zou, C.: SketchyCOCO: Image generation from freehand scene sketches. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
6. Ge, J., Cao, J., Zhu, X., Zhang, X., Liu, C., Wang, K., Liu, B.: Consistencies are all you need for semi-supervised vision-language tracking. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 1895–1904 (2024)
7. Gu, J., Meng, X., Lu, G., et al.: Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In: Advances in Neural Information Processing Systems. vol. 35, pp. 26418–26431 (2022)
8. Guo, Z., Wang, T.J.J., Pehlivan, S., Radman, A., Laaksonen, J.: PiTL: Cross-modal retrieval with weakly-supervised vision-language pre-training via prompting. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2261–2265 (2023)
9. He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., Neubig, G.: Towards a unified view of parameter-efficient transfer learning. In: International Conference on Learning Representations (2022)
10. Li, W., Qi, F., Sun, M., Yi, X., Zhang, J.: CCPM: A chinese classical poetry matching dataset. arXiv preprint arXiv:2106.01979 (2021)
11. Li, Y., Zhang, Y., Wu, B., Wen, J.R., Song, R., Bai, T.: A multi-modal knowledge graph for classical Chinese poetry. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 2318–2326 (2022)
12. Liu, S., Johns, E., Davison, A.J.: End-to-end multi-task learning with attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
13. Liu, Y., Cheng, M.M., Hu, X., Bian, J.W., Zhang, L., Bai, X., Tang, J.: Richer convolutional features for edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(8), 1939–1946 (2019)
14. Shan, B., Yin, W., Sun, Y., et al.: ERNIE-ViL 2.0: Multi-view contrastive learning for image-text pre-training. CoRR **abs/2209.15270** (2022)
15. Xie, C., Cai, H., Li, J., et al.: CCMB: A large-scale chinese cross-modal benchmark. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 4219–4227 (2023)
16. Yang, A., Pan, J., Lin, J., et al.: Chinese CLIP: Contrastive vision-language pretraining in chinese. arXiv preprint arXiv:2211.01335 (2022)
17. Zhang, R., Zhang, W., Fang, R., et al.: Tip-Adapter: Training-Free adaption of CLIP for Few-Shot classification. In: Computer Vision – ECCV. pp. 493–510 (2022)
18. Zhou, M., Yu, L., Singh, A., et al.: Unsupervised vision-and-language pretraining via retrieval-based multi-granular alignment. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16464–16473 (2022)