Human-Compatible CV

Weiyuan Ding[*] **Xiang Lian**[*]
Department of Electrical Engineering and Computer Sciences
Peking University
2000012991@stu.pku.edu.cn
2000012979@stu.pku.edu.cn

Abstract

Visual impairment is one of the major diseases affecting the quality of human life today. Among the many inconveniences caused by visual impairment, reading books with illustrations is an aspect worth noting. Due to the different layout and styles of illustrations, how to accurately extract illustrations and obtain appropriate descriptions is a difficult problem. In this paper, we design a complete pipeline to translate scanned books and generate complete text with illustration description information, aiming to help the visually impaired have a better reading experience. The relevant code will be published in *https://github.com/TheFatBlue/Human-compatibleCV.git*.

1 Introduction

Book reading is critical for enriching the spiritual world of the visually impaired. However, although the related technologies of CV are relatively mature, there are very few applications in this field. Therefore, it is essential to design such a pipeline to translate the scan copy of a book to a text with description.

The book translation pipeline can be roughly divided into three parts: book OCR and image extraction, Image Captioning and text combination. The technologies involved in mainly include OCR, Image extraction and IC(Image Captioning). As shown in Fig. 1, the complexity of book layout and painting style poses serious challenges to these three:

> **(1) Image Extraction:** A page may contain multiple pictures, and may also contain some small pictures that do not need to be paid attention to. How to extract pictures accurately and properly will test the robustness of the extraction algorithm.
>
> **(2) Noise Filtering:** There is a lot of noise in the result of direct OCR, including header and footer, pinyin, etc. These are things we don't want to see in the end. How to accurately filter out these noises needs to be considered.
>
> **(3) Image Captioning:** Illustration styles vary widely, including watercolors, abstract paintings, drawings, and more. Moreover, since the existing IC data sets[18, 32] are basically based on real pictures, some anthropomorphic styles in the illustrations will also cause great confusion to the model. Low resolution is also an issue to consider. Taking the above points together, generating an appropriate and accurate description is a challenging task.

In order to solve the above problems, we designed a highly robust and style-sensitive pipeline (Fig. 2) for book translation tasks to process the given ten books, and finally obtained smooth and relatively accurate results.

In addition, since the book descriptions manually marked are relatively simple, we proposed a method to supplement the manual annotations with the help of pictures to further improve the quality of the descriptions.

[*] Equal contribution.

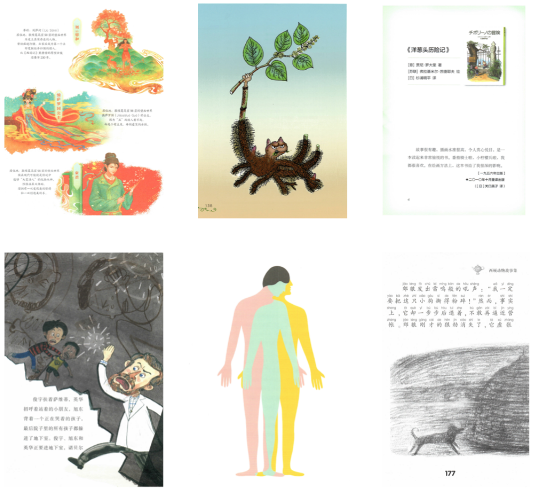


Figure 1: **Part of book typesetting and illustration styles.**

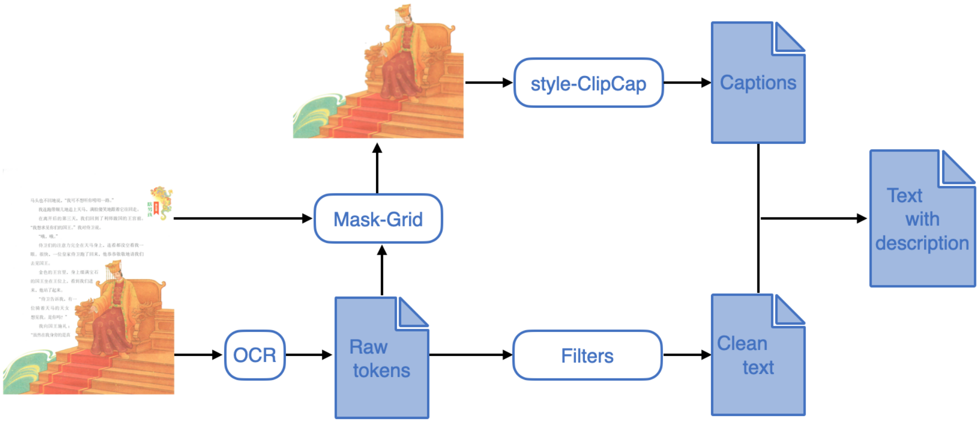


Figure 2: **Illustration for the translate pipeline.**

Our contributions are summarized as follows:

- We design a pipeline to translate scanned books into text with image descriptions.
- We design a pipeline to extract the pictures in the book and the descriptions in the corresponding manual annotations, and generate a dataset that can be used for training IC networks.
- We propose a conception about description supplementation, and provide a solution for implementation and evaluation.

2 Related Work

2.1 Image Captioning

Image Captioning(IC) tasks take an image as input and corresponding descriptions as output. The network encodes the image pixels to feature vectors and decodes to generate a sequence of words. Early works[3, 5] on IC usually use encoder from a pre-trained object detection network[25], which works well on COCO benchmark[18]. Later, the attention and transformer mechanism[28] allow network to focus on more specific visual features. More and more works[13] try to use a visual transformer as an encoder. Our work uses the expressive embedding of CLIP[22] for visual representation because of its training based on large dataset.

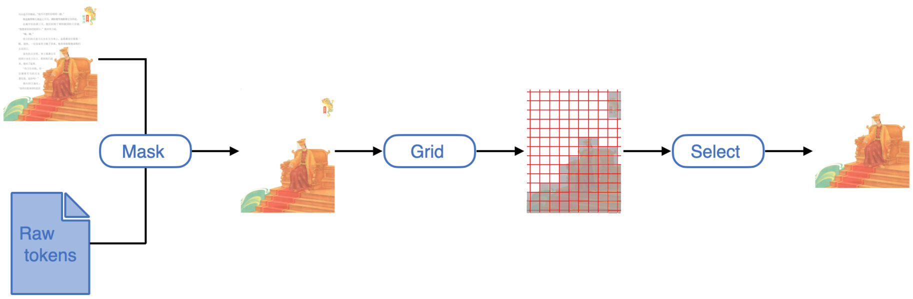


Figure 3: **Illustration for the mask-grid algorithm.**

For decoder, former works use LSTM varients[6, 30], while recent works[33] build on the transformer and use novel mechanism such as BERT[8]. Some visual information resides on the prefix, then a excellent auto-regressive model GPT-2[23] can be used.

Recent years, some solutions to image captioning on Chinese[16, 31] are proposed. However, the lack of dataset prevents the training. Some of the works try to use translated Flickr30k[32], but don't perform as well as those on English because of the language gap.

Also, some works on cross-domain style transfer are proposed. Part of them utilize the strong power of generative adversarial networks(GAN)[11], where most approaches[12] force the discriminator to learn to distinguish different styles form different domains and then use the adversarial loss from discriminator to supervise generators. Other works include Text style translation [27]which proposes to encode different styles into the same latent space and variational auto-encoders (VAE)[15].

2.2 CLIP Works

As a novel method, CLIP[22] jointly aggregate image and text descriptions. It was trained over more than 400 million image-text pairs, resulting in rich semantic latent space shared by both visual and textual data. A number of works on vision-language works[2, 10, 21] have achieve greater results based on the model.

3 Method

Since the task of book translation has its own uniqueness, such as the diversification of image layout and style, we propose a series of specific methods to solve these problems.

3.1 Data Generation

This part consists of extracting image-description pairs from the raw data. The image extraction method can also be applied to the pipeline of book translation.

3.1.1 Image Extraction

As shown in Fig. 3, we propose a non-machine learning, efficient and effective image extraction method called mask-grid. The proposal of this method mainly takes into account the characteristics that books usually have a white background: we first use the result of OCR[9] to mask the text part with white blocks, and then divide the whole page into several grids of fixed size. For each grid, if its average pixel value is lower than a threshold, then we consider that the grid contains a picture. For a grid containing a picture, if there are enough grids connected, then we think there is a picture here. In this way, we can not only get a white background, clean illustration, but also obtain more accurate position information than the simple bounding box —— this can effectively improve the vividness of the final text.

Since we set two thresholds here: the average pixel value and the number of grids, we can improve the robustness of the method very well: neither scanning noise nor small images that should not be considered will not affect the final result. And, although the complexity of this algorithm is $O(n^2)$, in fact this n will not be very large, so the process can be completed quickly.

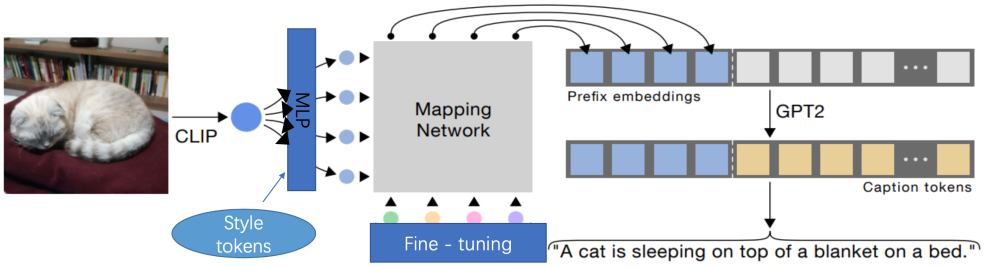


Figure 4: **Illustration for the modified model of ClipCap.** Image encoder: the CLIP model is used to encode the input image to obtain a picture vector clip_embed. Style blender: merge the CLIP prefix with style tokens. Mapping Network: acting as a bridge between the image space and the text space, it is responsible for mapping the image vector clip_embed into the text space, and obtaining a text prompt vector sequence prefix_embeds. Text decoder: use the GPT2 model to generate captions according to the prompt vector sequence prefix_embeds.

3.1.2 Description Extraction

Due to the diversity in the styles of manual annotations, it is difficult to accurately extract image descriptions from them. Finally, we extract all sentences containing the Chinese character "TU" to guarantee that no description is left, and then manually screened.

3.1.3 Image-Caption Pairing

In general, each extracted description can find a corresponding picture, but not vice versa, so we adopt the method of matching pictures for descriptions.

Since manual annotation does not include page number information, we first compare the OCR results with manual annotation to determine the page number range corresponding to each description. When comparing text, we use *jieba* for word segmentation and cosine similarity for computing similarity. Then, within the range of page numbers corresponding to the description, we match the first unmatched image for the description——this strategy handles most cases. For a small number of mismatches——such as multiple pages with multiple images in a row—we manually edit them.

Finally, we replaced the names of people with specific references in the descriptions with corresponding universal descriptions to facilitate subsequent training. The part about the text in the picture in the description is also deleted, which corresponds to the mask-grid method mentioned earlier.

3.2 Image Captioning

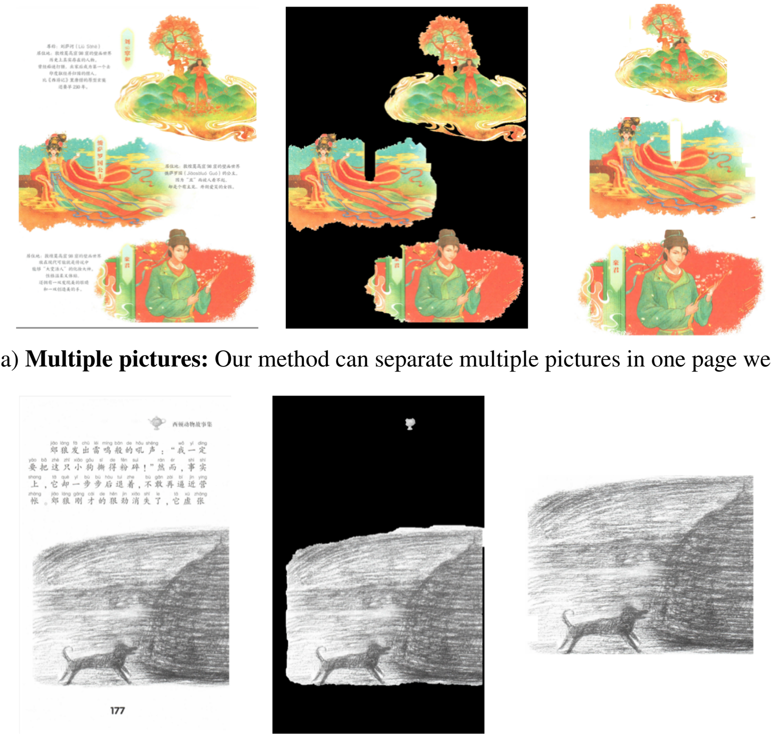
In this task, there are two main problems in IC: too few datasets, and style transfer is too difficult. We define this problem as the IC problem of small-sample style transfer, which is also the main consideration for our method selection.

Based on the ability of the Clip model to represent unknown categories, we choose ClipCap[19] as our baseline, and then translate the results into Chinese. As shown in Fig. 4, the model can be roughly divided into four parts : Clip encoder, style blender, mapping network and GPT-2[23] text decoder. For the consideration of training cost, we use MLP as the mapping network.

In order to enhance the model's ability to recognize different image styles, we added an MLP to the model to mix style token and the embeddings obtained by CLIP. The style token is a one-dimensional one-hot vector used to mark the style of the image. For the classification of styles, we simply consider the images of each book to belong to a separate category.

3.3 Text Relocation

This step is responsible for connecting the text recognized by OCR with the description generated by the image in a suitable way. First, the OCR results should be divided into semantically compliant sentences. We took a naive but effective approach: divide by periods. Although this method theoretically misses some divisions, its robustness to complex sentences is impressive in practice. Then we insert the descriptions corresponding to all the pictures on each page together with some conjunctions into the text on this page to get the final text.



(a) **Multiple pictures:** Our method can separate multiple pictures in one page well.

(b) **Small pictures:** Our method can ignore small pictures that should not be pay attention to.

Figure 5: **The comparison between ours and GrabCut.** (From left to right are the original image, GrabCut results, and our results.)

3.4 Description Supplement

Due to the limitation of time and computing power, we finally failed to implement a supplementary network, but made some adjustments on the basis of the current translation pipeline to complete it. The specific method is to first use the IC network to describe the picture, and then remove the part that is repeated with the manual description and stitch it together.

4 Experiments

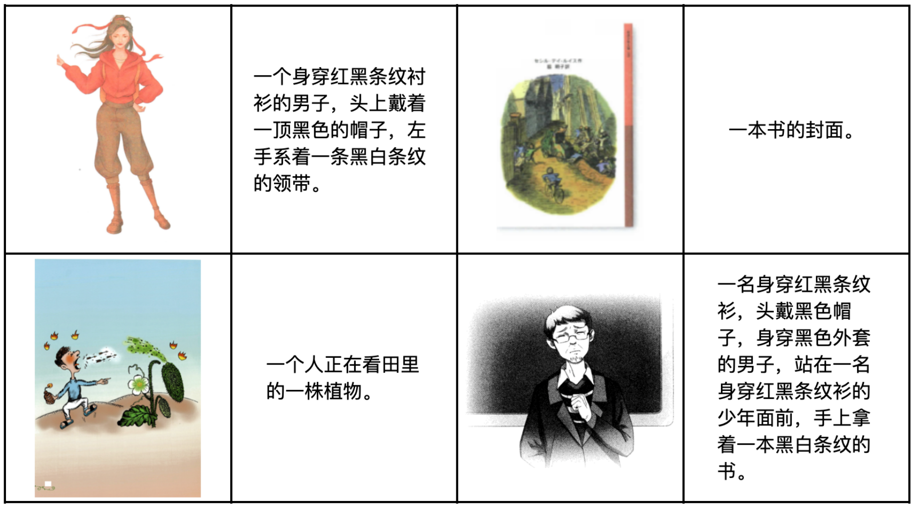
Dataset We end up extracting only about 400 image-description pairs from the given ten books. A data set of this size is far from enough, so we decided to adopt a fine-tuning method based on the model trained with the existing data set. We tried the Flickr30k[32] and MS COCO[4] datasets separately, and finally found that the latter performed relatively better.

Image Extraction Here we can compare the original image with the results of GrabCut[26] and mask-grid (Fig. 5) (the three rightmost images in Fig. 5a are separated). It can be found that mask-grid can better extract each picture separately from multiple pictures, and has good robustness to noise such as small pictures.

Image Captioning We adopted different kinds of Mapping Networks and training strategies, used our data set for fine-tuning on the pre-trained model, and evaluated the generation effect on the test set according to the evaluation method of the COCO data set. We use common matrices BLEU[20], CIDEr[29], METEOR[7], ROGUE[17] and SPICE[1]. The results are shown in Tab. 1. Among them, Modified is our improved network with style token added. The one marked with $*$ is the result of fine-tuning the original model first, and adding the style token to continue fine-tuning on this basis. Due to the slow convergence, we trained for a total of 60 (10+50) epochs. It can be seen that no matter which Mapping Network is adopted, the representation ability of the model can be significantly improved after adding the style token.

Model	Mapping Network		epoch		B@4 ↑	CIDEr ↑	METEOR ↑	ROGUE ↑	SPICE ↑
	MLP	Transformer	10	20					
ClipCap	✓		✓		0.06	0.06	0.06	0.14	0.08
	✓			✓	0.08	0.09	0.06	0.17	0.08
		✓	✓		0.13	0.10	0.07	0.17	0.05
		✓		✓	0.17	0.09	0.08	0.19	0.10
Modified	✓		✓		0.16	0.11	0.07	0.16	0.07
	✓			✓	0.18	0.16	0.08	**0.21**	**0.12**
	✓			✓*	**0.21**	0.10	**0.09**	0.19	0.09
		✓	✓		0.12	0.09	**0.09**	0.17	0.06
		✓		✓	0.20	**0.18**	**0.09**	**0.21**	**0.12**

Table 1: **Quantitative evaluation.** It can be seen that no matter which Mapping Network is adopted, the representation ability of the model can be significantly improved after adding the style token.



一个身穿红黑条纹衬衫的男子，头上戴着一顶黑色的帽子，左手系着一条黑白条纹的领带。

一本书的封面。

一个人正在看田里的一株植物。

一名身穿红黑条纹衫，头戴黑色帽子，身穿黑色外套的男子，站在一名身穿红黑条纹衫的少年面前，手上拿着一本黑白条纹的书。

Figure 6: **Some results of the image captioning.**

Fig. 6 shows some of the IC results. It can be seen that the model can describe the content in the pictures relatively accurately, but it has a certain lack of ability to describe the details.

Description Supplement Based on the task, we propose an evaluation method for the result, which contains two parts. **(1) Similarity with the original description**: Using simple methods for determining semantic relevance such as BLEU and cosine similarity. **(2) Validity of the added description**: With DALLE[24] as a generator, we get corresponding images based on supplementary descriptions and compare their similarity with the original images using FID[14]. Fig. 7 shows an example that works relatively well. This method can reflect the effect of supplementary description to a certain extent, but it also relies heavily on the generator. If the style of the generated image differs greatly from the original image, the reference value of the similarity is low.

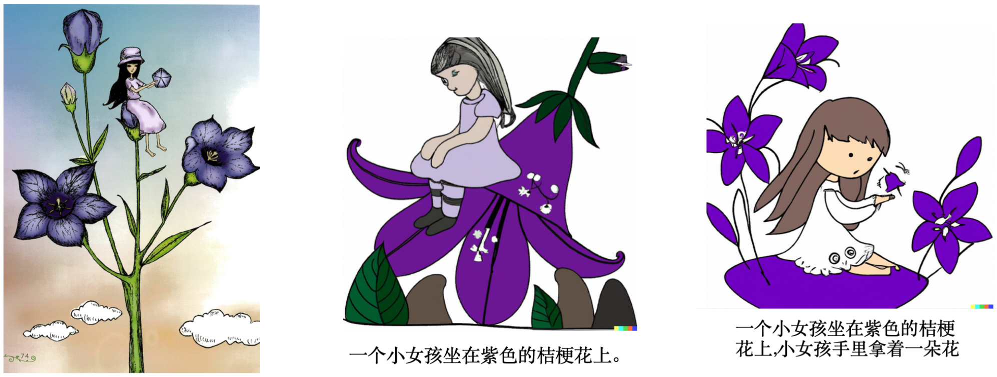
5 Discussion

There are still some deficiencies in our current work, if possible, we will continue to improve in the future.

(i) At present, the data generation part still needs more manual corrections. Perhaps using a CNN network to assist in matching can further enhance the robustness of the algorithm and reduce the workload of manual corrections.

(ii) Even with the addition of the style token, the cross-domain capability of the model still needs to be improved. A more robust network can be considered to alleviate this problem.

(iii) When evaluating supplementary results, consider selecting a better generator, or provide certain guidance for it to ensure the reference value of its results.


一个小女孩坐在紫色的桔梗花上。
一个小女孩坐在紫色的桔梗花上,小女孩手里拿着一朵花

(a) **Original picture**. (b) **Manual description**. (c) **Supplement description**.

Figure 7: **A example of supplement.** When evaluating similarity, loss in manual description is 335.52, and 318.10 in supplementary description.

(iv) Design a network to supplement the manual description. The initial idea is: extract the subject from the description, then go to the picture to find the corresponding subject and extract the details to add to the description. Its core is to find a better feature representation.

6 Conclusion

We generated an IC dataset with multiple styles to train a style-sensitive IC network, and applied it to our designed book translation pipeline, which finally automatically converted 10 scanned books into text with illustration descriptions . We also propose a method that complements human descriptions and evaluate the results. For further work, we plan to continue to optimize our pipeline to better accommodate various book styles

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, 2016. 5

[2] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *ArXiv*, abs/2103.10951, 2021. 3

[3] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6298–6306, 2016. 2

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. 5

[5] Xinlei Chen and C. Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. *ArXiv*, abs/1411.5654, 2014. 2

[6] Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, and W. Liu. Regularizing rnns for caption generation by reconstructing the past with the present. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7995–8003, 2018. 3

[7] Michael J. Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *WMT@ACL*, 2014. 5

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019. 3

[9] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Hongya Wang. Pp-ocr: A practical ultra lightweight ocr system. *ArXiv*, abs/2009.09941, 2020. 3

[10] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ArXiv*, abs/2108.00946, 2021. 3

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 3

[12] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. Mscap: Multi-style image captioning with unpaired stylized text. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4199–4208, 2019. 3

[13] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *Neural Information Processing Systems*, 2019. 2

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 6

[15] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, H. Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *ArXiv*, abs/1512.09300, 2015. 3

[16] Dong J et al Li X, Lan W. Adding chinese captions to images. 2016. 3

[17] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain, July 2004. 5

[18] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 1, 2

[19] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 4

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 2002. 5

[21] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2065–2074, 2021. 3

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2, 3

[23] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 3, 4

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 6

[25] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015. 2

[26] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004. 5

[27] Tianxiao Shen, Tao Lei, Regina Barzilay, and T. Jaakkola. Style transfer from non-parallel text by cross-alignment. *ArXiv*, abs/1705.09655, 2017. 3

[28] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017. 2

[29] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2014. 5

[30] Oriol Vinyals, Alexander Toshev, Samy Bengio, and D. Erhan. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2014. 3

[31] WANG Mingwen et al XIAO Yuhan, JIANG Aiwen. Chinese image captioning based on middle-level visual-semantic composite attributes. *Journal of Chinese information*, 2019. 3

[32] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1, 3, 5

[33] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *ArXiv*, abs/2101.00529, 2021. 3

A Appendix

A.1 Authors' Contribution

Weiyuan Ding: image extraction, pipeline implement, network modification, experiment running

Xiang Lian: text extraction, data generation, network modification, evaluation, experiment running

A.2 Acknowledgements

Thanks very much for the acknowledgements of teachers and TA since the beginning of the semester. Through this course, both theoretical and practical aspects of computer vision have a more in-depth field. I feel very good at listening to the class, and I can obviously feel the teachers and TA's intentions in the homework design. Finally, I sincerely wish the course better and better. Happy Lunar New Year for teachers and TA!