Supplementary Material of
A Simple Framework for Text-Supervised Semantic Segmentation

A. Implementation Details

A.1. Implementation Details of Training

The implementation follows ZeroVL [3], which provides a training guidance that allows to conduct CLIP [12] with less resource and public data.

Architecture. The image and text encoders are ViT-S/16 [6] and BERT-base [5], respectively. The image and text encoders are trained from scratch. The encoded image and text features are projected into 512-dim embeddings by single layer perceptrons.

Training. AdamW [10] optimizer is used for training, and the weight decay is 0.2. The learning rate is initialized to 3e-4. The learning rate schedule is cosine decay, with a minimum learning rate at 3e-5 (*i.e.*, minimum scale of 0.1). The model is trained for 20 epochs, where warmup epochs account for 2.5% of the entire training procedure. The temperature of contrastive loss is learnable and initialized to 0.07. The batch size is 4096.

config	value
optimizer	AdamW
base learning rate	3e-4
weight decay	0.2
optimizer momentum	β_1, β_2=0.9, 0.98
batch size	4096
learning rate schedule	cosine decay
minimum lr scale	0.1
training epochs	20
warmup proportion	0.025
base temperature	0.07

Table 5. **Training setting.**

Augmentation. During pre-training, batches are comprised by randomly sampling image-text pairs from pre-training datasets. We apply two augmentation techniques for images, the first one is "crop and resize", the second one is AutoAugment [2].

The "crop and resize" operation comes from [3,12], each image is randomly cropped to a rectangular region with aspect ratio sampled in $[3/4, 4/3]$. The ratio of preserved area is sampled in $[60\%, 100\%]$. Then we resize the cropped region to 224×224 resolution.

AutoAugment [2] searches data augmentation policies with reinforcement learning and considers a wide range of operations including translation, rotation, shearing, color normalization, *etc*. We adopt the AutoAugment policy learned on ImageNet.

Regarding the text modality, 20% input words are processed during augmentation. For each word, we mask it, replace it with a random word, or delete it with a probability of 50%, 10% and 40%, respectively.

A.2. Implementation Details of Evaluation

Augmentation. No augmentation is used during zero-shot semantic segmentation evaluation except for resizing the images to 288 × 288 resolution.

COCO protocol. We evaluate the performance of our method in the COCO-Stuff dataset [1]. Following the practice of [13], 80 foreground object classes are used. We combine the instance masks of the same category to get the semantic segmentation mask for each image.

DenseCRF parameters. We adopt the two most-common pairwise potentials (*i.e.* Gaussian and bilateral), with the default arguments. The inference has 3 iterations.

B. Additional Experiments

We first examine the effects of LoDA on zero-shot image-text retrieval and linear probing classification tasks. Next, we reveal some ablation details.

B.1. Image-Text Retrieval

To evaluate the representation ability of our proposed LoDA method, Flickr30K [11] and MSCOCO [9] benchmarks are leveraged for zero-shot image-text retrieval tasks.

As reported in Table 6 and 7, our proposed LoDA method outperforms the baseline in every metric. Regarding RSUM, our method is 1.5 points higher than baseline on Flickr30K and 8.3 points highter on MSCOCO. Zero-shot image-text retrieval performance is an important prerequisite for semantic segmentation. The retrieval ability helps to provide correct object classes of images.

method	Flickr30K						
	image-to-text			text-to-image			
	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
w/o LoDA	78.7	94.0	97.3	60.8	84.9	90.5	506.2
w/ LoDA	78.8	94.3	97.5	61.2	85.0	90.9	507.7

Table 6. The LoDA ablation results of zero-shot image-text retrieval on Flickr30K datasets.

method	MSCOCO						
	image-to-text			text-to-image			
	R@1	R@5	R@10	R@1	R@5	R@10	RSUM
w/o LoDA	49.9	75.7	83.8	34.4	60.9	71.6	376.3
w/ LoDA	51.8	76.6	85.4	35.7	62.3	72.8	384.6

Table 7. The LoDA ablation results of zero-shot image-text retrieval on MSCOCO datasets.

B.2. Zero-shot Classification

Evaluation of the linear classification task is performed on ImageNet [4] dataset. As shown in table 8, LoDA improves the zero-shot classification performance.

method	Acc@1	Acc@5
w/o LoDA	46.6	74.2
w/ LoDA	47.1	74.6

Table 8. The LoDA ablation results of zero-shot classification on ImageNet-1k.

B.3. Linear Classification

config	value
optimizer	LARS
base learning rate	0.1
optimizer momentum	0.9
batch size	32768
learning rate schedule	cosine decay
warmup epochs	10
training epochs	90

Table 9. **Linear probing settings.**

Evaluation of the linear classification task is performed on ImageNet [4] dataset. We follow the setup in MAE [7] to evaluate linear classification performance. The optimizer is LARS [14] without weight decay. The base learning rate is 0.1. We train with batch size 32768 for 90 epochs, where warmup epochs are 10. For data augmentations, we perform standard cropping and resizing. The classifier is a single-layer perceptron.

As shown in Table 10, our method achieves 0.6 points higher than baseline on top-1 accuracy and 0.2 points higher on top-5 accuracy. These results signify the effectiveness and robustness of the proposed LoDA approach.

method	Acc@1	Acc@5
w/o LoDA	64.6	87.3
w/ LoDA	65.2	87.5

Table 10. LoDA ablation of linear classification on ImageNet-1k.

B.4. Ablation Details

In Section 5.3, we study the most significant hyper-parameters $\mathcal{M}^I$ and $\mathcal{M}^T$ in evaluation and pre-training phases. We use figures rather than tables for a more vivid illustration. The detailed mIoU results of these ablation experiments are reported in Table 11 and 12. Best results are **bold** and default settings are marked in gray.

mIoU		$\mathcal{M}^I$		
		1	3	5
$\mathcal{M}^T$	1	55.4	56.3	**56.6**
	3	45.1	45.2	45.9
	5	33.3	33.0	33.0

Table 11. Evaluations on PASCAL VOC with various $\mathcal{M}^I$ and $\mathcal{M}^T$ in zero-shot evaluation.

mIoU		$\mathcal{M}^I$		
		1	3	5
$\mathcal{M}^T$	1	40.9	48.7	37.3
	3	54.1	55.3	51.9
	5	54.3	56.3	**56.6**

Table 12. Evaluations on PASCAL VOC with various $\mathcal{M}^I$ and $\mathcal{M}^T$ in pre-training.

C. Visualizations

C.1. Similarity Maps

We verified that LoDA addresses the problems of dense alignment. *i.e.*, (1) *LoDA makes vision encoder perceives main objects.* (2) *LoDA makes main objects and context equally significant in the image-text contrasting.* Similar with Figure 3, we provide more examples in Figure 10.

C.2. Segmentation Results

PASCAL VOC 2012. We illustrate in Figure 11 the qualitative results of our approach on the PASCAL VOC 2012 *validation* set. Our approach correctly identifies target objects and produces accurate masks.

PASCAL Context. We show additional qualitative results of our approach on the PASCAL Context *validation* set.

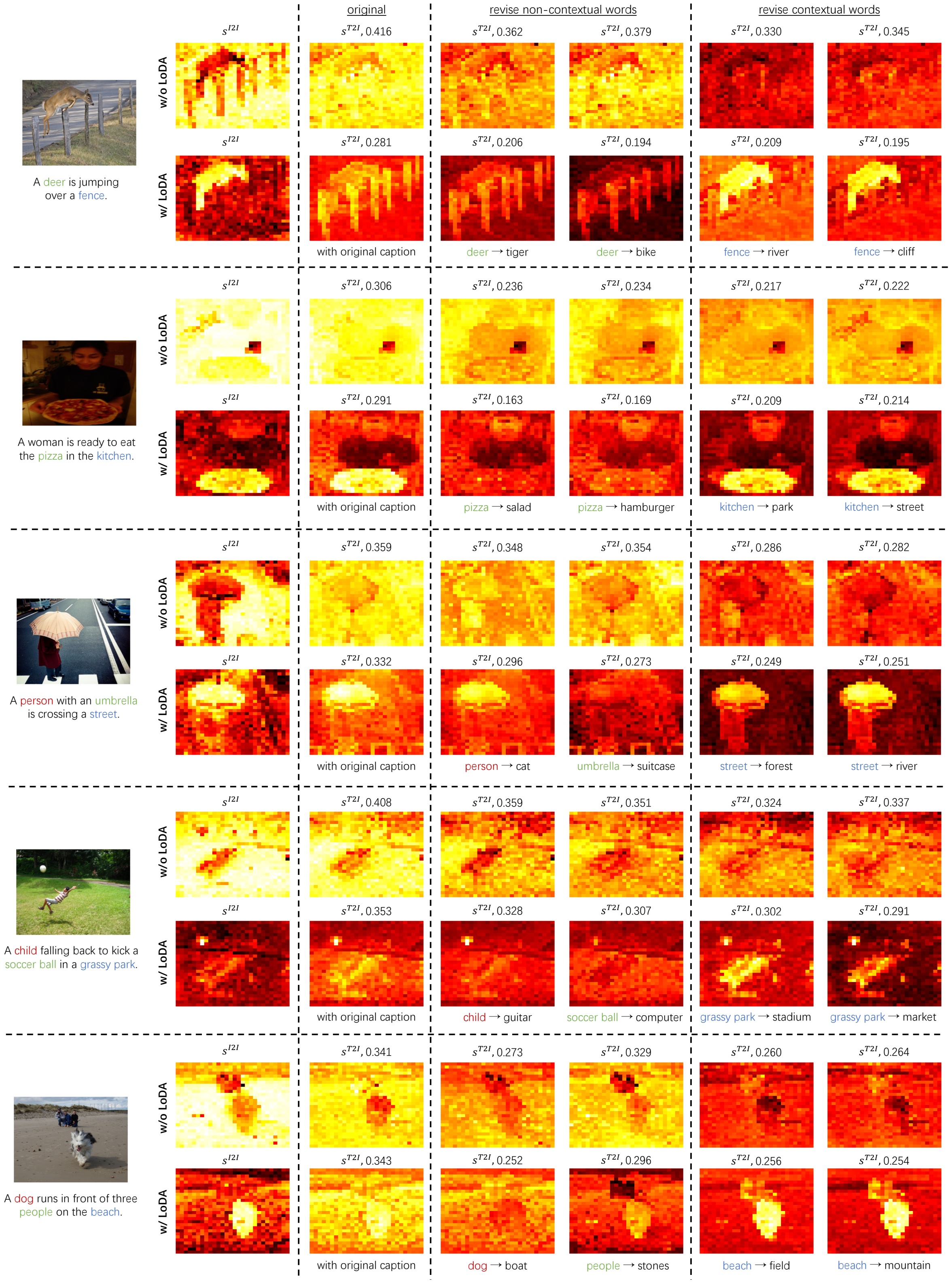

Figure 10. Visualization of more patch-wise similarity maps on Flickr30K *test* set. For each sample, we show (1) original image-text pair, (2) s^{I2I}, (3) s^{T2I} regarding to the original caption, and (4) s^{T2I} regarding to manually revised captions (non-contextual words *vs.* contextual words). In each revised caption, the modified key entity words are marked in colors. For s^{T2I} maps, the overall image-text similarity score provided by CLIP is attached.

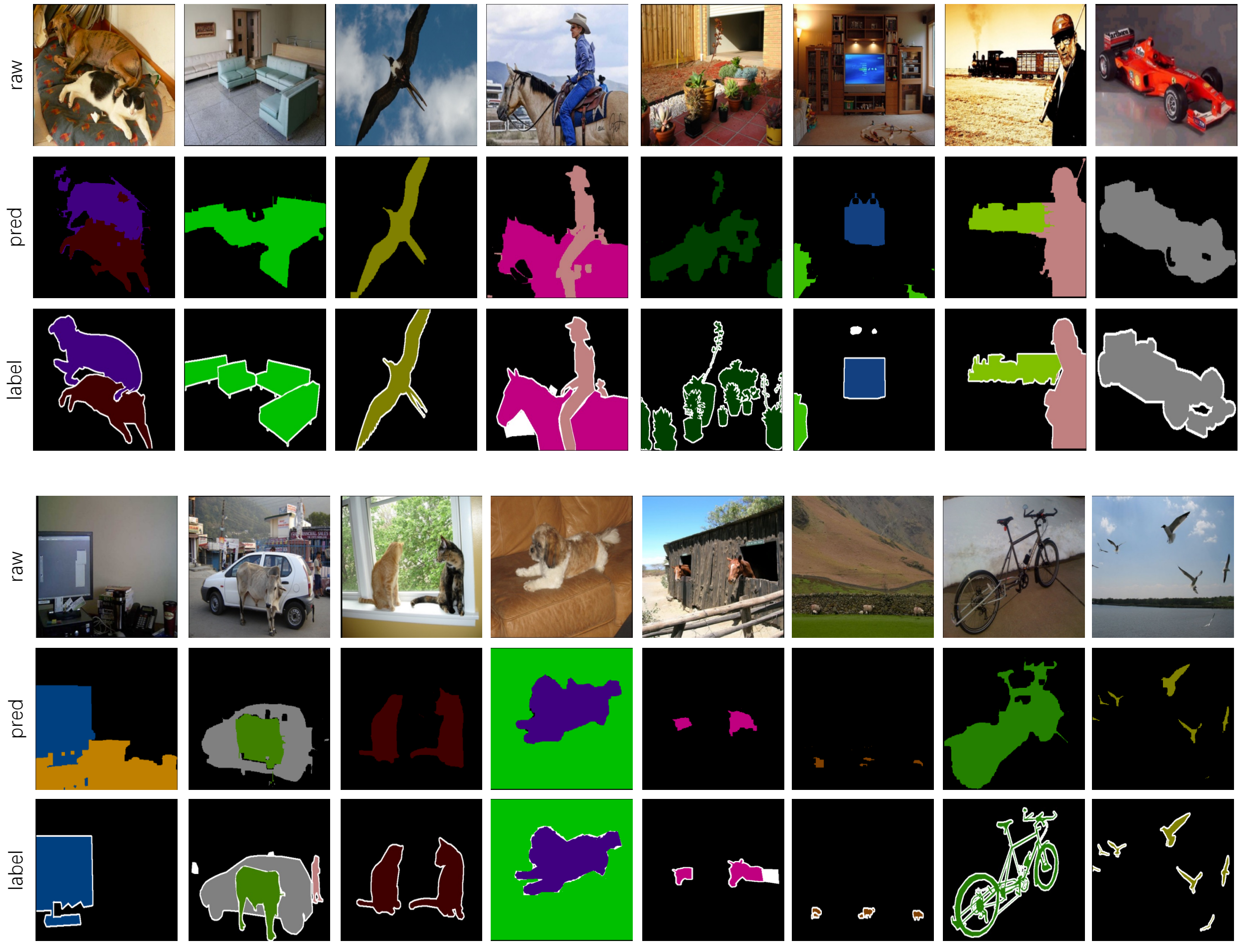

Figure 11. Qualitative results of SimSeg on PASCAL VOC.

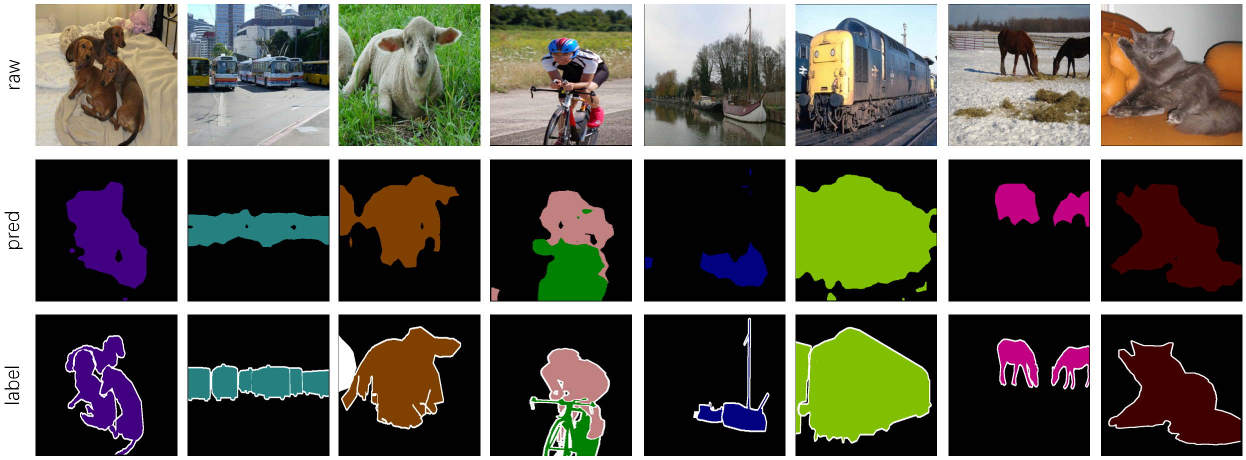

Figure 12. Qualitative results of SimSeg (w/o CRF) on PASCAL VOC.

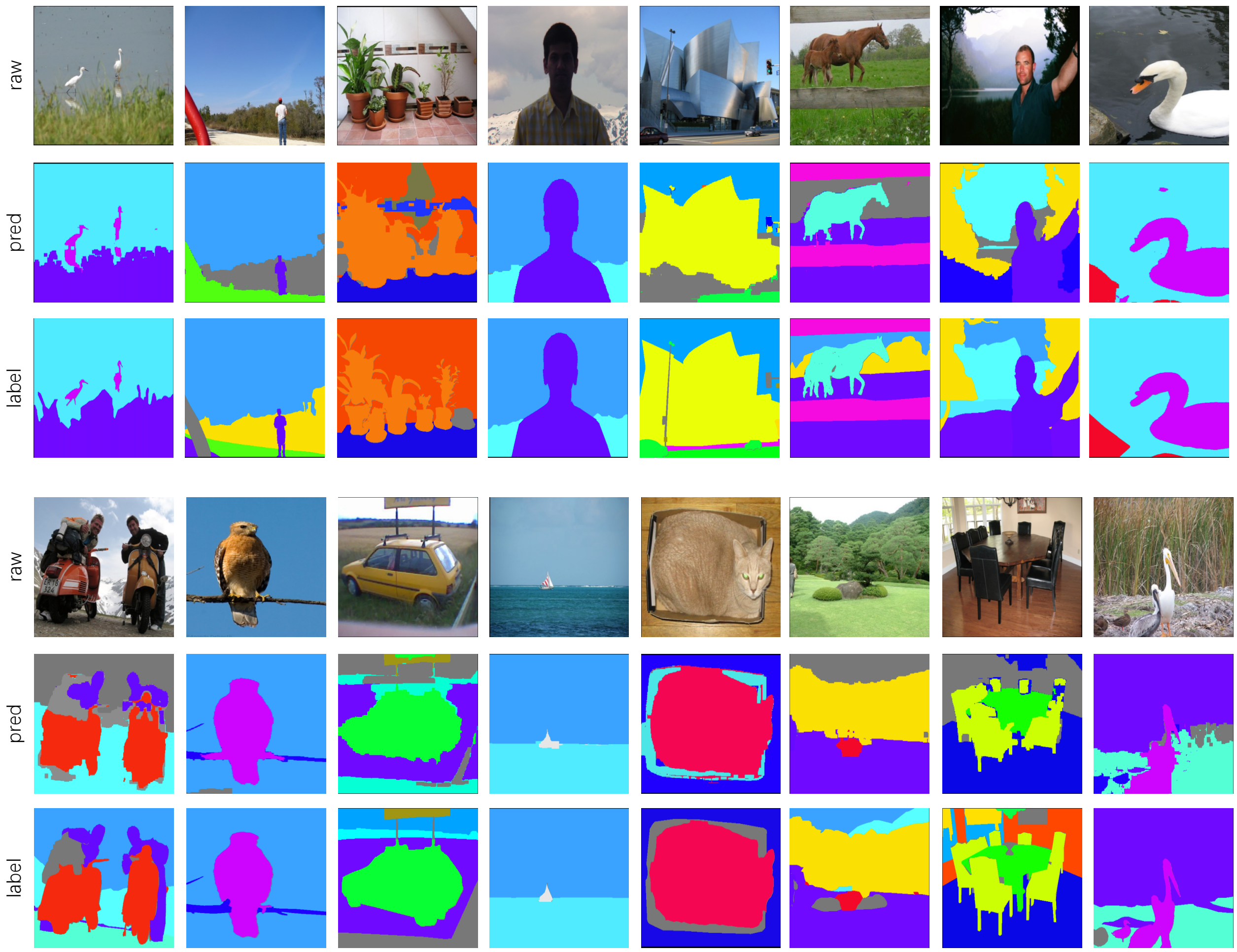

Figure 13. Qualitative results of SimSeg on PASCAL Context.

C.3. Effects of CRF

SimSeg does *not* rely on CRF [8] to generate fine masks. Qualitative results of the “w/o CRF” setting are shown in Figure 12.

C.4. Effects of High-Frequency Entities

Our method is inferior to predict high-frequency entities. In web image-text datasets, “person” is one of the most high-frequency entities. We visualize several bad cases in Figure 14. When person co-exists with other objects, the model cannot produce segmentation on person. It could result from the confidence scores of “person” are greatly lower than other categories. For instance, the image-text score of “person” class is ∼0.15, while “horse” is ∼0.25. Our model fails to select “person” in the thresholding post-processing. It indicates a drawback of CLIP-driven zero-shot semantic segmentation, *i.e.*, high-frequency entities cannot play an important role in image-text contrasting, resulting in low image-text similarity score on classes concerning high-frequency entities.

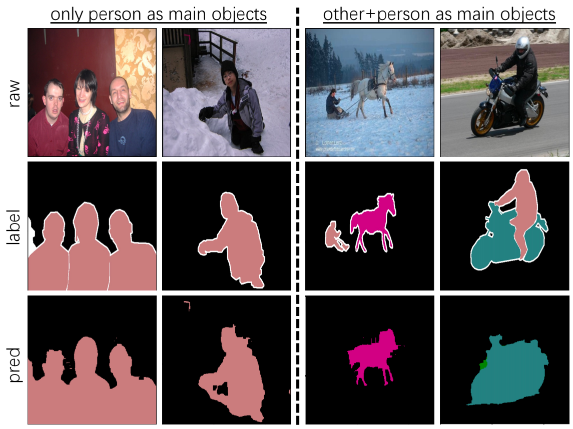

Figure 14. Bad cases on the high-frequency entity “person”.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, 2018.

[2] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. AutoAugment: Learning augmentation strategies from data. In *CVPR*, 2019.

[3] Quan Cui, Boyan Zhou, Yu Guo, Weidong Yin, Hao Wu, Osamu Yoshie, and Yubo Chen. Contrastive vision-language pre-training with limited resources. In *ECCV*, 2022.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[8] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[11] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, J. Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 2015.

[12] 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 *ICML*, 2021.

[13] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *CVPR*, 2022.

[14] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv:1708.03888*, 2017.