CLAMP: Prompt-based Contrastive Learning for Connecting Language and Animal Pose
(Supplementary Material)

Xu Zhang[1] Wen Wang[2] Zhe Chen[1] Yufei Xu[1] Jing Zhang[1] Dacheng Tao[1]

[1]The University of Sydney, Australia [2]Zhejiang University, China

{xzha0930,yuxu7116}@uni.sydney.edu.au wwen@zju.edu.cn

{zhe.chen1,jing.zhang1}@sydney.edu.au dacheng.tao@gmail.com

1. Appendix

In the supplementary material, we show the complexity analysis of our CLAMP and add another ablation study to study the impact of the regularization from losses. Then we evaluate the generalization ability of the proposed CLAMP model by involving more unseen animal species in the zero-shot learning setting. In addition, we visualize the spatial-level score map to further validate the effectiveness of the spatial-level adaptation process in CLAMP by explicitly demonstrating the established spatial connections. At last, we show some pose estimation results from SimpleBaseline [2] and CLAMP for qualitative comparison, which shows that CLAMP can perform robustly on different animal species with different postures and sizes.

1.1. Complexity Analysis

We performed the complexity analysis of our CLAMP and the SimpleBaseline method, and the results are in Table S1. The numbers of parameters and GFLOPs are calculated on 256 by 256 images. The results show that our CLAMP brings significant accuracy improvements with a little complexity increase. In practice, the increased complexity is mainly from the cross-attention and prompt encoder. Since our proposed adaptation schemes mainly improve training, they only bring less than 0.01 extra GFLOPs and 0 extra parameters during inference. Note that the prompt embeddings encoded by the text encoder can be stored offline after training and are shared by all the test images, so the text encoder will not add complexity to inference.

1.2. Additional Ablation Study

We validate that our method can effectively leverage CLIP's capability with the following results. We tested the models' performance under the setting where the E_{prompt} is replaced by a trainable matrix for training, obtaining a performance gain of merely 0.2 AP *w.r.t.* the baseline. Alternatively, our proposed method achieved a gain of 1.7 AP with the help of feature-aware and spatial-aware adaptation. This demonstrates that the major improvements come from our adaptation schemes rather than other components like initialization and regularization from losses. The results are shown in Table S2.

Method	Backbone	Params(M)	GFLOPs	AP
SimpleBaseline [2]	ResNet-50	49	9.0	70.9
CLAMP (ours)	ResNet-50	68	9.2	72.9
SimpleBaseline [2]	ViT-Base	91	16.6	72.6
CLAMP (ours)	ViT-Base	98	16.8	74.3

Table S1. Complexity comparison.

Method	Backbone	Embedding type	AP
SimpleBaseline [2]	ViT-Base	w/o	72.6
CLAMP (ours)	ViT-Base	trainable matrix	72.8
CLAMP (ours)	ViT-Base	prompt embedding	74.3

Table S2. Ablation study of regularization from losses.

We also conduct an ablation study for the ViT-Large backbone which has a large number of parameters, and the results can be found in Table S3. Even for the significantly large model such as ViT-Large, CLAMP is still effective in taking advantage of the language knowledge for animal pose estimation.

1.3. Additional Zero-shot Experimental Results

We report more zero-shot learning experimental results in addition to the results in the main paper to validate the models' generalization ability on unseen animal species, *i.e.*, a) training the model using animals from Bovidae and testing the model using instances from Equidae and Felidae, and b) training the model using animals from Canidae and testing the model using instances from Cricetidae and Equidae. The same training settings as described in the zero-shot experimental setting in the main text are adopted

Method	Backbone	Pre-train	AP	AP_{50}	AP_{75}	AP_M	AP_L	AR
SimpleBaseline [2]	ViT-Large	CLIP	76.9	96.0	84.4	56.5	77.2	80.0
CLAMP (ours)	ViT-Large	CLIP	77.8	96.8	85.0	58.7	78.1	81.0

Table S3. Ablation study for ViT-Large on AP-10K [3].

Method	Backbone	Train	Test	AP	AP_{50}	AP_{75}	AP_M	AP_L	AR
SimpleBaseline [2]	ResNet-50	Bovidae	Equidae	41.9	71.8	40.3	27.3	42.0	46.6
CLAMP (ours)	ResNet-50	Bovidae	Equidae	46.6	75.6	47.5	48.3	46.6	51.2
SimpleBaseline [2]	ResNet-50	Bovidae	Felidae	22.0	52.4	15.0	11.2	22.1	28.4
CLAMP (ours)	ResNet-50	Bovidae	Felidae	28.7	67.6	18.9	11.9	29.0	36.3
SimpleBaseline [2]	ResNet-50	Canidae	Cricetidae	16.1	41.7	10.1	3.4	16.5	26.0
CLAMP (ours)	ResNet-50	Canidae	Cricetidae	22.0	51.1	14.1	10.1	22.7	31.6
SimpleBaseline [2]	ResNet-50	Canidae	Equidae	20.5	43.9	16.6	11.1	20.7	25.3
CLAMP (ours)	ResNet-50	Canidae	Equidae	28.4	59.1	23.1	9.4	29.0	34.1

Table S4. Additional comparisons of the zero-shot generalization performance of different methods on AP-10K [3].

and the results are available in Table S4. We can observe that with the help of pose-specific text prompts and the decomposed adaptation process, our CLAMP outperforms the SimpleBaseline by a large margin, *e.g.*, 46.6 AP vs. 41.9 AP, 28.7 AP vs. 22.0 AP, 22.0 AP vs. 16.1 AP, and 28.4 AP vs. 20.5 AP, respectively in these four settings. Such observation validates that language knowledge can improve the model's generalization ability since the shared language knowledge of keypoints can alleviate the difficulties caused by large visual inter- and intra-species variances.

1.4. Performance on the human pose estimation dataset

Our method can be extended to human pose by replacing the $KeyPoint$ in Eq.(2) in the main text with the names of human keypoints, but we found that the human pose estimation methods can easily achieve compelling results by taking advantage of rich labeled data. This can marginalize the benefits of the knowledge in CLIP. In practice, we have tested our method on the COCO human pose dataset [1], obtaining 0.4 AP improvement. However, we found that our CLAMP can bring more benefits for human pose estimation in low-data regimes. We test our method on random $k\%$ of COCO data for $k \in \{1, 2, 3, 5, 10\}$ and display the mean results of three times random sampling on COCO in Table. S5.

1.5. Visualization of Spatial-level Score Maps

Based on Eq.(3) and Eq.(4) in the main text, we can obtain the keypoint presence score on each spatial position of the input image with the help of spatial-level loss $\mathcal{L}_{spatial}$. We visualize the upsampled score maps in Fig. S1, which displays the established spatial connections between language descriptions and image features in our CLAMP. It can be seen that for each keypoint description, the highest score values show up in the corresponding image region that has the same semantics as the keypoint description. This indicates that the spatial-level loss helps establish spatial connections between language knowledge and visual features, which can provide positional information for animal poses.

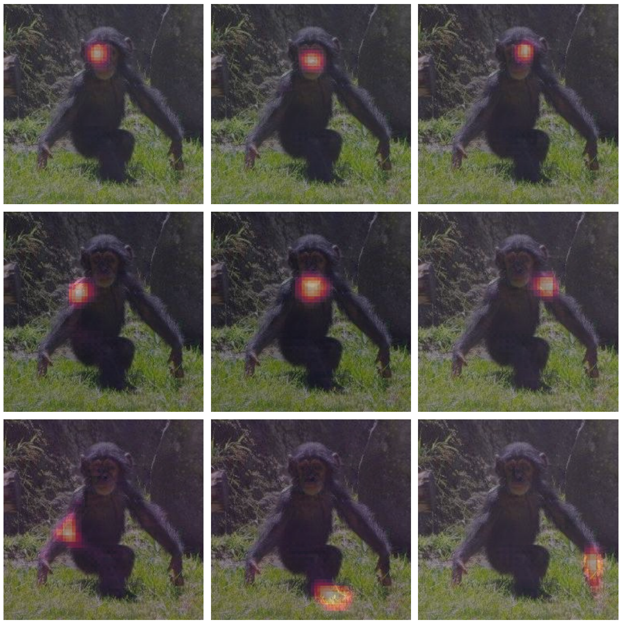

Figure S1. Visualization of the spatial-level score maps. In order to intuitively display the relationship between the keypoint presence score and the animal image, we superimpose the obtained score maps on the animal image. As shown in the figure, the lightness and darkness of the pixels are used to indicate the keypoint presence score (the brighter pixels indicate the higher scores). The superimposed images for different keypoints are displayed in the following order: right eye, nose, left eye, right shoulder, neck, left shoulder, right elbow, right front paw, and left front paw (from the first row to the last row, from left to right in each row).

Method	Backbone	Pre-train	$AP@1\%$	$AP@2\%$	$AP@3\%$	$AP@5\%$	$AP@10\%$
SimpleBaseline [2]	ViT-Base	CLIP	49.5	53.3	55.5	58.4	62.1
CLAMP (ours)	ViT-Base	CLIP	53.1	56.8	57.6	59.5	62.7

Table S5. Performance on COCO [1] in low-data regimes. Note that $AP@k\%$ means the AP on random $k\%$ of COCO data.

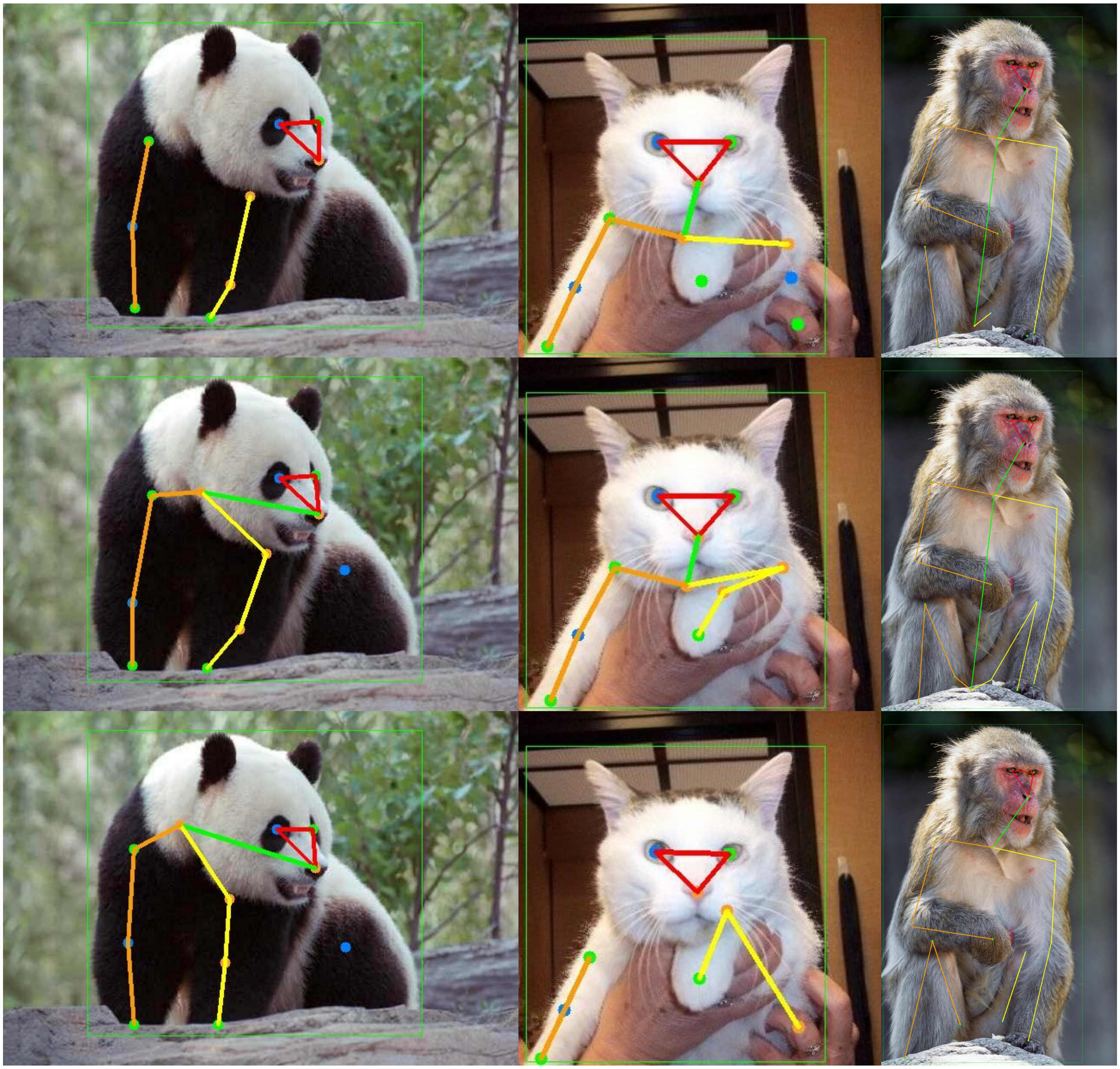

Figure S2. Qualitative analysis of the SimpleBaseline [2] (the first row) and the proposed CLAMP (the second row). The ground truth poses are shown in the last row.

1.6. Qualitative Analysis

To get an intuitive understanding of the proposed method, we show some qualitative results in Fig. S2 and Fig. S3. In each figure, the baseline method, *i.e.*, SimpleBaseline [2], the proposed CLAMP, and the ground truth are shown from top to bottom. As can be seen, our method can produce accurate pose estimation results on animals with large variances in appearances and poses. Taking the first

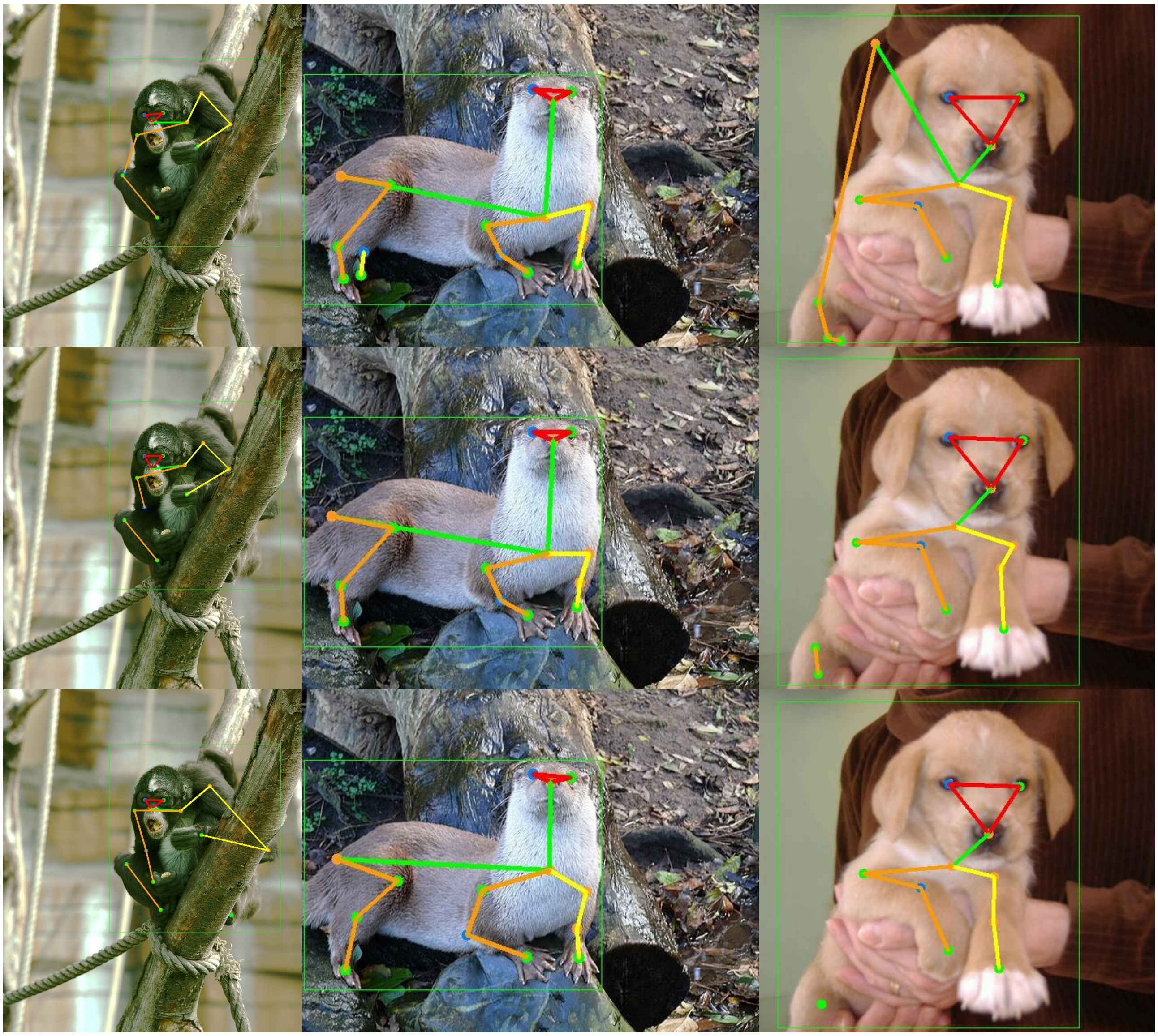

Figure S3. More qualitative results of the SimpleBaseline [2] (the first row) and the proposed CLAMP (the second row). The ground truth poses are shown in the last row.

column in Fig. S2 as an example, the baseline method in the first row overlooks the neck and left hip of the panda. By contrast, our CLAMP successfully leverages the language knowledge and outputs all keypoints that are labeled in the ground truth. Similar results can also be observed in the images of other animal species. Besides, CLAMP can sometimes produce more accurate results than human annotations. For example, in the second column, the neck, the left elbow, and the left shoulder are neglected or probably incorrectly labeled. By contrast, our CLAMP can locate and recognize those keypoints, demonstrating its potential in dealing with hard cases.

References

[1] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 3

[2] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 1, 2, 3, 4

[3] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. *arXiv preprint arXiv:2108.12617*, 2021. 2