Label Propagation for Zero-shot Classification with Vision-Language Models -Supplementary-

Vladan Stojnić[1] Yannis Kalantidis[2] Giorgos Tolias[1]

[1] VRG, FEE, Czech Technical University in Prague [2] NAVER LABS Europe

Impact of hyper-parameters We show the effect of γ and α in Table 1 and observe stability over a wide range of values for both cases.

Additional backbones In Table 2, we present transductive and inductive zero-shot classification results on ImageNet with additional CLIP backbones. We present results with two versions of ViT-L-14 from CLIP [4]. Additionally, we present the results with ViT-B-16 and ViT-H-14 from OpenCLIP [1][1] trained on the LAION-2B [5] dataset. We see from Table 2 that ZLaP improves the results with different backbones in both transductive and inductive setups. This verifies that ZLaP is not backbone dependant and that it is independent of the dataset used for pre-training.

Per-dataset results In Tables 3 and 4, we present per dataset results for transductive and inductive setups, respectively.

Leveraging LLM generated prompts In the main paper, we present the average results when we leverage LLM generated prompts from CuPL [2]. In Tables 5 and 6, we present per dataset results for transductive and inductive setups, respectively. CuPL prompts, compared to handcrafted universal class templates, improve CLIP+ZLaP from 60.0% to 64.6% and from 58.7% to 64.2% for the transductive and inductive setup, respectively.

[1] https://github.com/mlfoundations/open_clip

γ \ α	0.01	0.05	0.1	0.3	0.5	0.7	0.8	0.9
0.5	33.6	34.4	35.2	37.9	39.9	42.8	43.9	43.6
1	37.5	38.3	39.2	41.6	43.8	46.0	46.8	45.9
5	50.9	51.3	51.5	52.1	52.5	52.3	51.4	49.7
8	52.2	52.7	52.9	53.7	53.6	52.6	51.5	49.6
10	52.8	52.9	53.2	53.8	53.6	52.4	51.4	49.2
12	52.8	52.8	53.1	53.6	53.3	52.4	51.0	49.0
15	1.5	52.0	52.0	53.0	52.8	51.7	50.4	48.6
20	0.5	0.5	0.5	27.0	48.1	47.5	46.2	44.9

Table 1. **Impact of γ and α hyper-parameters.** Results presented on CUB dataset for transductive inference.

	Transductive	Inductive
Results with ViT-L-14		
CLIP	75.9	75.9
+ **ZLaP**	**77.2**	**77.3**
Results with ViT-L-14@336		
CLIP	77.0	77.0
+ **ZLaP**	**78.0**	**78.4**
Results with ViT-B-16 (LAION-2B)		
CLIP	70.4	70.4
+ **ZLaP**	**72.0**	**72.1**
Results with ViT-H-14 (LAION-2B)		
CLIP	78.0	78.0
+ **ZLaP**	**79.1**	**79.1**

Table 2. **Accuracy on ImageNet using different CLIP backbones.**

Web-crawled unlabeled images We construct a new set of unlabeled images with 10,000 images per class that are chosen either randomly, or based on proximity of their image or text features to the class representation. Results are presented in Table 7. Switching to using only the LAION-based unlabeled set, we observe that random selection fails by performing worse than CLIP, but the other two options provide some improvement, with the caption-based neighbors being a bit better. Interestingly, web-crawling is better than the the target distribution images for the Pets dataset, while much worse for Eurosat due to the lack of satellite images on LAION. On the other hand, if the randomly selected set is mixed with that from the target distribution, ZLaP manages to benefit from the relevant images and to deliver an improvement compared to CLIP.

	imagenet	dtd	eurosat	fgvca	flowers	food	pets	sun	cars	caltech	cifar10	cifar100	cub	ucf	avg
Results with ResNet50															
TPT†	60.7	40.8	28.3	17.6	62.7	74.9	84.5	61.5	58.5	87.0	–	–	–	**69.8**	*(58.8)*
CLIP-DN	60.2	41.1	28.4	17.3	63.3	77.2	83.1	60.9	54.8	88.3	74.0	44.7	48.9	60.4	57.3
CLIP	60.3	41.1	26.9	16.7	62.9	76.6	83.1	61.2	54.4	87.9	72.3	42.5	47.0	59.9	56.6
+ **ZLaP**	61.8	41.9	**35.5**	17.8	65.9	78.8	83.9	63.3	57.8	**89.6**	78.2	47.9	52.1	65.9	60.0
(vs CLIP)	↑1.5	↑0.8	↑8.6	↑1.1	↑3.0	↑2.2	↑0.8	↑2.1	↑3.4	↑1.7	↑5.9	↑5.4	↑5.1	↑6.0	↑3.4
InMaP	**63.8**	44.8	33.4	**19.0**	65.0	**79.4**	89.0	65.3	61.5	74.5	78.9	49.6	**55.5**	65.6	60.4
+ **ZLaP**	**63.8**	**45.9**	34.5	18.4	**67.1**	79.2	**89.2**	**65.9**	**62.0**	80.7	**79.2**	**49.7**	55.3	67.8	**61.3**
(vs CLIP)	↑3.5	↑4.8	↑7.6	↑1.7	↑4.2	↑2.6	↑6.1	↑4.7	↑7.6	↓7.2	↑6.9	↑7.2	↑8.3	↑7.9	↑4.7
(vs InMaP)	↑0.0	↑1.1	↑1.1	↓0.6	↑2.1	↓0.2	↑0.2	↑0.6	↑0.5	↑6.2	↑0.3	↑0.1	↓0.2	↑2.2	↑0.9
Results with ViT-B/16															
TPT†	69.0	47.8	42.4	24.8	69.0	84.7	87.8	65.5	66.9	**94.2**	–	–	–	68.0	*(65.5)*
CLIP-DN	68.3	45.7	53.3	24.3	68.0	86.0	87.7	66.5	64.0	93.6	91.4	69.6	56.1	68.4	67.3
CLIP	68.8	45.1	50.2	23.0	67.0	85.7	88.3	66.3	63.8	93.9	91.2	68.7	55.2	67.5	66.8
+ **ZLaP**	69.7	46.0	57.7	26.3	67.9	87.2	87.9	67.8	66.8	91.8	92.6	70.8	58.2	73.8	68.9
(vs CLIP)	↑0.9	↑0.9	↑7.5	↑3.3	↑0.9	↑1.5	↓0.4	↑1.5	↑3.0	↓2.1	↑1.4	↑2.1	↑3.0	↑6.3	↑2.1
InMaP	72.5	50.9	60.1	28.3	70.8	**88.0**	**93.2**	71.3	71.7	76.7	93.3	**73.3**	63.8	75.7	70.7
+ **ZLaP**	**72.7**	**51.8**	**60.9**	**28.4**	**73.4**	87.9	92.8	**71.9**	**72.1**	83.7	**93.6**	**73.3**	**64.1**	**77.7**	**71.7**
(vs CLIP)	↑3.9	↑6.7	↑10.7	↑5.4	↑6.4	↑2.2	↑4.5	↑5.6	↑8.3	↓10.2	↑2.4	↑4.6	↑8.9	↑10.2	↑4.9
(vs InMaP)	↑0.2	↑0.9	↑0.8	↑0.1	↑2.6	↓0.1	↓0.4	↑0.6	↑0.4	↑7.0	↑0.3	↑0.0	↑0.3	↑2.0	↑1.0

Table 3. **Trasductive zero-shot classification accuracy on 14 datasets** for two CLIP backbones. Rows denoted as *(vs CLIP)* and *(vs InMaP)* show the absolute accuracy gains of our method over CLIP and InMaP, respectively. † denotes numbers taken from InMaP [3].

	imagenet	dtd	eurosat	fgvca	flowers	food	pets	sun	cars	caltech	cifar10	cifar100	cub	ucf	avg
Results with ResNet50															
TPT†	60.7	40.8	28.3	17.6	62.7	74.9	84.5	61.5	58.5	87.0	–	–	–	**69.8**	*(58.8)*
CLIP-DN	60.2	41.2	28.3	17.2	63.3	77.2	83.3	60.8	54.9	**88.3**	74.0	44.7	48.9	60.4	57.3
CLIP	60.3	41.1	26.9	16.7	62.9	76.7	83.1	61.2	54.5	87.9	72.3	42.5	47.0	59.9	56.6
+ **ZLaP**	62.2	42.8	31.9	17.4	69.3	77.9	80.3	61.8	56.4	86.9	76.3	46.0	49.7	62.8	58.7
(vs CLIP)	↑1.9	↑1.7	↑5.0	↑0.7	↑6.4	↑1.2	↓2.8	↑0.6	↑1.9	↓1.0	↑4.0	↑3.5	↑2.7	↑2.9	↑2.1
+ **ZLaP***	62.9	43.1	**38.8**	17.9	68.8	78.3	77.7	61.2	55.8	86.3	78.6	48.0	51.1	64.2	59.5
(vs CLIP)	↑2.6	↑2.0	↑11.9	↑1.2	↑5.9	↑1.6	↓5.4	↑0.0	↑1.3	↓1.6	↑6.3	↑5.5	↑4.1	↑4.3	↑2.9
InMaP	62.9	45.7	33.6	**19.2**	66.4	**79.2**	**85.7**	65.0	**62.0**	76.2	79.0	49.7	**55.4**	66.0	60.4
+ **ZLaP**	62.9	**46.6**	36.3	18.7	69.1	79.0	83.4	64.9	61.8	80.2	79.1	**50.6**	54.9	66.5	**61.0**
(vs CLIP)	↑2.6	↑5.5	↑9.4	↑2.0	↑6.2	↑2.3	↑0.3	↑3.7	↑7.3	↓7.7	↑6.8	↑8.1	↑7.9	↑6.6	↑4.4
(vs InMaP)	↑0.0	↑0.9	↑2.7	↓0.5	↑2.7	↓0.2	↓2.3	↓0.1	↓0.2	↑4.0	↑0.1	↑0.9	↓0.5	↑0.5	↑0.6
+ **ZLaP***	**63.0**	46.3	36.2	18.9	**69.4**	**79.2**	81.4	**65.1**	61.9	79.3	**79.2**	50.5	55.1	67.0	60.9
(vs CLIP)	↑2.7	↑5.2	↑9.3	↑2.2	↑6.5	↑2.5	↓1.7	↑3.9	↑7.4	↓8.6	↑6.9	↑8.0	↑8.1	↑7.1	↑4.3
(vs InMaP)	↑0.1	↑0.6	↑2.6	↓0.3	↑3.0	↑0.0	↓4.3	↑0.1	↓0.1	↑3.1	↑0.2	↑0.8	↓0.3	↑1.0	↑0.5
Results with ViT-B/16															
TPT†	69.0	47.8	42.4	24.8	69.0	84.7	87.8	65.5	66.9	**94.2**	–	–	–	68.0	*(65.5)*
CLIP-DN	68.3	45.6	53.3	24.3	67.9	86.0	87.7	66.5	64.1	93.6	91.5	69.6	56.0	68.4	67.3
CLIP	68.8	45.1	50.2	23.0	67.0	85.7	88.3	66.3	63.8	93.9	91.2	68.7	55.2	67.5	66.8
+ **ZLaP**	70.2	48.6	55.6	25.4	73.5	86.9	87.1	67.4	65.6	93.1	92.2	71.0	59.4	71.5	69.1
(vs CLIP)	↑1.4	↑3.5	↑5.4	↑2.4	↑6.5	↑1.2	↓1.2	↑1.1	↑1.8	↓0.8	↑1.0	↑2.3	↑4.2	↑4.0	↑2.3
+ **ZLaP***	71.0	49.1	58.2	25.8	72.6	87.3	86.3	67.2	66.1	92.1	92.7	72.0	59.1	72.2	69.4
(vs CLIP)	↑2.2	↑4.0	↑8.0	↑2.8	↑5.6	↑1.6	↓2.0	↑0.9	↑2.3	↓1.8	↑1.5	↑3.3	↑3.9	↑4.7	↑2.6
InMaP	72.0	49.6	59.4	29.0	71.9	**87.9**	**91.6**	**71.4**	**71.9**	79.0	93.4	73.7	63.9	75.4	70.7
+ **ZLaP**	**72.1**	**51.2**	**63.2**	**29.1**	**75.9**	87.8	90.0	71.0	71.2	84.0	93.4	74.0	**64.3**	**76.3**	**71.7**
(vs CLIP)	↑3.3	↑6.1	↑13.0	↑6.1	↑8.9	↑2.1	↑1.7	↑4.7	↑7.4	↓9.9	↑2.2	↑5.3	↑9.1	↑8.8	↑4.9
(vs InMaP)	↑0.1	↑1.6	↑3.8	↑0.1	↑4.0	↓0.1	↓1.6	↓0.4	↓0.7	↑5.0	↑0.0	↑0.3	↑0.4	↑0.9	↑1.0
+ **ZLaP***	**72.1**	51.0	62.7	29.0	75.5	**87.9**	89.0	**71.4**	71.8	83.1	**93.6**	**74.2**	64.2	**76.3**	71.6
(vs CLIP)	↑3.3	↑5.9	↑12.5	↑6.0	↑8.5	↑2.2	↑0.7	↑5.1	↑8.0	↓10.8	↑2.4	↑5.5	↑9.0	↑8.8	↑4.8
(vs InMaP)	↑0.1	↑1.4	↑3.3	↑0.0	↑3.6	↑0.0	↓2.6	↑0.0	↓0.1	↑4.1	↑0.2	↑0.5	↑0.3	↑0.9	↑0.9

Table 4. **Inductive zero-shot classification accuracy on 14 datasets** for two CLIP backbones. Rows denoted as *(vs CLIP)* and *(vs InMaP)* show the absolute accuracy gains of our method over CLIP and InMaP, respectively. * denotes our method with approximation of $\hat{Y}$. † denotes numbers taken from InMaP [3].

	imagenet	dtd	fgvca	flowers	food	pets	sun	cars	caltech	cifar10	cifar100	ucf	avg
Results with ResNet50													
CLIP	61.7	49.1	18.5	67.9	77.8	87.5	63.8	55.8	88.7	76.4	45.2	63.5	63.0
+ **ZLaP**	62.7	51.4	20.2	67.6	78.9	88.1	65.2	58.8	**89.8**	77.6	47.4	67.8	64.6
(vs CLIP)	↑1.0	↑2.3	↑1.7	↓0.3	↑1.1	↑0.6	↑1.4	↑3.0	↑1.1	↑1.2	↑2.2	↑4.3	↑1.6
InMaP	**64.4**	54.5	**22.2**	67.2	**79.3**	**89.9**	67.4	62.8	73.7	78.2	50.2	68.2	64.8
+ **ZLaP**	64.3	**55.6**	**22.2**	**69.8**	79.2	89.5	**67.8**	**63.2**	78.9	**78.9**	**50.5**	**70.2**	**65.8**
(vs CLIP)	↑2.6	↑6.5	↑3.7	↑1.9	↑1.4	↑2.0	↑4.0	↑7.4	↓9.8	↑2.5	↑5.3	↑6.7	↑2.8
(vs InMaP)	↓0.1	↑1.1	↑0.0	↑2.6	↓0.1	↓0.4	↑0.4	↑0.4	↑5.2	↑0.7	↑0.3	↑2.0	↑1.0
Results with ViT-B/16													
CLIP	70.0	53.2	27.9	73.4	86.3	91.7	69.5	66.1	**94.4**	90.7	69.4	70.5	71.9
+ **ZLaP**	70.5	54.0	30.1	72.2	86.9	91.8	69.7	67.3	92.7	92.4	69.9	74.0	72.6
(vs CLIP)	↑0.5	↑0.8	↑2.2	↓1.2	↑0.6	↑0.1	↑0.2	↑1.2	↓1.7	↑1.7	↑0.5	↑3.5	↑0.7
InMaP	**73.3**	57.3	**31.9**	74.1	**88.1**	**93.7**	73.3	**72.8**	78.0	93.4	**73.3**	77.1	73.9
+ **ZLaP**	**73.3**	**57.9**	31.7	**76.9**	88.0	93.3	**73.7**	**72.8**	83.3	**93.6**	73.2	**79.5**	**74.8**
(vs CLIP)	↑3.3	↑4.7	↑3.8	↑3.5	↑1.7	↑1.6	↑4.2	↑6.7	↓11.1	↑2.9	↑3.8	↑9.0	↑2.9
(vs InMaP)	↑0.0	↑0.6	↓0.2	↑2.8	↓0.1	↓0.4	↑0.4	↑0.0	↑5.3	↑0.2	↓0.1	↑2.4	↑0.9

Table 5. **Transductive zero-shot classification accuracy on 12 datasets** for two CLIP backbones and prompts generated by a LLM [2]. Rows denoted as *(vs CLIP)* and *(vs InMaP)* show the absolute accuracy gains of our method over CLIP and InMaP, respectively.

	imagenet	dtd	fgvca	flowers	food	pets	sun	cars	caltech	cifar10	cifar100	ucf	avg
Results with ResNet50													
CLIP	61.7	49.1	18.5	67.9	77.8	**87.5**	63.8	55.8	**88.7**	76.4	45.2	63.5	63.0
+ **ZLaP**	63.1	51.4	20.0	**72.7**	78.4	85.4	63.3	57.8	88.3	77.9	48.0	63.6	64.2
(vs CLIP)	↑1.4	↑2.3	↑1.5	↑4.8	↑0.6	↓2.1	↓0.5	↑2.0	↓0.4	↑1.5	↑2.8	↑0.1	↑1.2
InMaP	**63.4**	**54.6**	**22.6**	68.8	**79.1**	86.4	**66.6**	**62.5**	75.7	78.2	50.4	67.5	64.6
+ **ZLaP**	**63.4**	54.1	**22.6**	71.5	**79.1**	83.4	66.5	62.4	79.4	**78.8**	**51.0**	**67.7**	**65.0**
(vs CLIP)	↑1.7	↑5.0	↑4.1	↑3.6	↑1.3	↓4.1	↑2.7	↑6.6	↓9.3	↑2.4	↑5.8	↑4.2	↑2.0
(vs InMaP)	↑0.0	↓0.5	↑0.0	↑2.7	↑0.0	↓3.0	↓0.1	↓0.1	↑3.7	↑0.6	↑0.6	↑0.2	↑0.4
Results with ViT-B/16													
CLIP	70.0	53.2	27.9	73.4	86.3	91.7	69.5	66.1	**94.4**	90.7	69.4	70.5	71.9
+ **ZLaP**	71.2	55.5	29.8	77.7	87.2	91.1	69.7	67.5	**94.4**	91.6	71.3	72.6	73.3
(vs CLIP)	↑1.2	↑2.3	↑1.9	↑4.3	↑0.9	↓0.6	↑0.2	↑1.4	↑0.0	↑0.9	↑1.9	↑2.1	↑1.4
InMaP	72.4	**57.2**	**32.8**	75.8	**88.0**	**92.3**	**73.0**	**72.9**	79.8	93.3	73.7	**76.6**	74.0
+ **ZLaP**	**72.5**	56.3	**32.8**	**78.5**	87.9	89.6	72.6	**72.9**	83.9	**93.5**	**73.8**	**76.6**	**74.2**
(vs CLIP)	↑2.5	↑3.1	↑4.9	↑5.1	↑1.6	↓2.1	↑3.1	↑6.8	↓10.5	↑2.8	↑4.4	↑6.1	↑2.3
(vs InMaP)	↑0.1	↓0.9	↑0.0	↑2.7	↓0.1	↓2.7	↓0.4	↑0.0	↑4.1	↑0.2	↑0.1	↑0.0	↑0.2

Table 6. **Inductive zero-shot classification accuracy on 12 datasets** for two CLIP backbones and prompts generated by a LLM [2]. Rows denoted as *(vs CLIP)* and *(vs InMaP)* show the absolute accuracy gains of our method over CLIP and InMaP, respectively.

	imagenet	dtd	eurosat	fgvca	flowers	food	pets	sun	cars	caltech	cifar10	cifar100	cub	ucf	avg.
Results with ResNet50															
CLIP	60.3	41.1	26.9	16.7	62.9	76.7	83.1	61.2	54.5	87.9	72.3	42.5	47.0	59.9	56.6
+ **ZLaP** (target distribution)	**62.2**	**42.8**	**31.9**	**17.4**	**69.3**	**77.9**	80.3	61.8	**56.4**	86.9	**76.3**	**46.0**	**49.7**	**62.8**	**58.7**
+ **ZLaP** (target distr. + LAION random)	61.4	42.3	30.2	15.7	63.5	77.1	80.3	61.6	53.7	87.8	75.1	42.9	47.5	59.8	57.1
+ **ZLaP** (LAION random)	59.9	41.4	26.2	14.3	59.1	74.5	79.3	61.1	51.4	87.6	70.6	41.4	43.4	58.8	54.6
+ **ZLaP** (LAION image neighbors)	60.6	41.1	29.1	16.7	63.5	76.9	83.5	61.9	54.7	**88.4**	69.5	41.1	48.2	59.6	56.8
+ **ZLaP** (LAION caption neighbors)	60.7	40.5	26.9	16.9	63.0	76.9	**83.6**	**62.0**	55.3	**88.4**	73.0	41.7	48.5	60.1	57.0
Results with ViT-B/16															
CLIP	68.8	45.1	50.2	23.0	67.0	85.7	88.3	66.3	63.8	93.9	91.2	68.7	55.2	67.5	66.8
+ **ZLaP** (target distribution)	**70.2**	**48.6**	**55.6**	**25.4**	**73.5**	**86.9**	87.1	**67.4**	**65.6**	93.1	**92.2**	**71.0**	**59.4**	**71.5**	**69.1**
+ **ZLaP** (target distr. + LAION random)	69.5	45.9	53.1	21.0	67.3	86.3	86.4	66.9	64.7	93.7	91.9	69.3	55.6	67.6	67.1
+ **ZLaP** (LAION random)	68.6	44.9	49.4	19.8	65.3	85.3	86.8	66.5	63.0	93.6	90.3	68.6	54.0	66.9	65.9
+ **ZLaP** (LAION image neighbors)	69.0	45.4	49.2	23.8	68.1	85.8	**88.4**	66.9	64.5	**94.2**	90.8	68.1	56.7	67.6	67.0
+ **ZLaP** (LAION caption neighbors)	69.1	45.0	49.4	23.4	68.1	85.9	**88.4**	67.0	64.6	94.0	90.8	68.5	56.7	67.6	67.1

Table 7. **Inductive zero-shot classification accuracy on 14 datasets using different sources of unlabeled data.** Compared to the original experiments that use unlabeled images from the target distribution, LAION-400M is used to create a web-crawled unlabeled set.

References

[1] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 1

[2] Sarah M. Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, 2023. 1, 4

[3] Qi Qian, Yuanhong Xu, and Juhua Hu. Intra-modal proxy learning for zero-shot visual categorization with clip. In *NeurIPS*, 2023. 2, 3

[4] 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. 1

[5] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 1