Supplementary material

1 Implementation Details

The resolution of input images is resized to 224x224 for both pre-training on training set and evaluation on downstream tasks. The maximum sequence length of the tokens is limited to 77. We only use the center crop for data augmentation in the training. The minimal and maximal scores $s_{\min} = 0$, $s_{\max} = 0.7$ used in the margin loss, and the threshold $\omega = 0.05$. For the weight norm clipping, we set γ to 0.7 on CC2M dataset. The code is released at https://github.com/Rhyssiyan/IncCLIP.PyTorch

2 Prompt Templates

Table 2 demonstrates the prompt templates used for downstream image classification datasets. We replace "{}" with label name during inference. We also report the prompt templates used for image-text retrieval task in Table 4.

3 More results on Conceptual Caption dataset

CI-CC2M dataset is the Conceptual Caption subset used for class incremental split in main body, which is built by giving a pseudo label to each image using the ImageNet pretrained classification model and picking the 2M image-text pairs that contain all 1K classes. We establish a 4-chunks instance incremental split on the CI-CC2M dataset by randomly splitting it into 4 chunks with 0.5M image-text pairs per chunk for the sake of experiment completion. The results of zero-shot image classification and image-text retrieval tasks are shown in Table 1 and Table 3, respectively. Our method is consistently better than other comparison methods. We will release the splits in the future.

Table 2: Prompt Templates for downstream image classification datasets

Dataset	Template

ImageNet	a photo of a {}, a bad photo of a {}, a photo of many {}, a sculpture of a {}, a photo of the hard to see {}, a low resolution photo of the {}, a rendering of a {}, graffiti of a {}, a bad photo of the {}, a cropped photo of the {}, a tattoo of a {}, the embroidered {}, a photo of a hard to see {}, a bright photo of a {}, a photo of a clean {}, a photo of a dirty {}, a dark photo of the {}, a drawing of a {}, a photo of my {}, the plastic {}, a photo of the cool {}, a close-up photo of a {}, a black and white photo of the {}, a painting of the {}, a painting of a {}, a pixelated photo of the {}, a sculpture of the {}, a bright photo of the {}, a cropped photo of a {}, a plastic {}, a photo of the dirty {}, a jpeg corrupted photo of a {}, a blurry photo of the {}, a photo of the {}, a good photo of the {}, a rendering of the {}, a {} in a video game, a photo of one {}, a doodle of a {}, a close-up photo of the {}, the origami {}, the {} in a video game, a sketch of a {}, a doodle of the {}, a origami {}, a low resolution photo of a {}, the toy {}, a rendition of the {}, a photo of the clean {}, a photo of a large {}, a rendition of a {}, a photo of a nice {}, a photo of a weird {}, a blurry photo of a {}, a cartoon {}, art of a {}, a sketch of the {}, a embroidered {}, a pixelated photo of a {}, itap of the {}, a jpeg corrupted photo of the {}, a good photo of a {}, a plushie {}, a photo of the nice {}, a photo of the small {}, a photo of the weird {}, the cartoon {}, art of the {}, a drawing of the {}, a photo of the large {}, a black and white photo of a {}, the plushie {}, a dark photo of a {}, itap of a {}, graffiti of the {}, a toy {}, itap of my {}, a photo of a cool {}, a photo of a small {}, a tattoo of the {}
CIFAR10	a rendition of a {}, a drawing of the {}, the {} in a video game, a painting of the {}, a sculpture of a {}, art of the {}, a photo of many {}, a rendering of a {}, a tattoo of the {}, a rendition of the {}, a blurry photo of a {}, a drawing of a {}, a painting of a {}, graffiti of the {}, a pixelated photo of a {}, a rendering of the {}, a sketch of a {},a cropped photo of a {}, a sketch of the {}, the embroidered {}, a photo of a hard to see {}, a jpeg corrupted photo of a {}, a tattoo of a {}, a black and white photo of a {}, a good photo of a {}
CIFAR100	a rendition of a {}, a doodle of a {}, a photo of the cool {}, a sketch of a {}, a good photo of a {}, a bright photo of the {}, the {} in a video game, a painting of a {}, a pixelated photo of a {}, a pixelated photo of the {}, a photo of many {}, a painting of the {}, a photo of a small {}, a sketch of the {}, graffiti of a {}, a close-up photo of a {}, a embroidered {}, art of a {}, a drawing of the {}
SUN397	a photo of a {}
Food101	a photo of a {}, a type of food {}.
Flowers102	'a photo of a {}, a type of flower {}', 'a flower photo of a {}.

DTD	a photo of a {} texture
Caltech101	a doodle of a {}, a rendering of a {}, a good photo of a {}, art of a {}

References

1.
2. Cha, H., Lee, J., Shin, J.: Co2l: Contrastive continual learning. In: CVPR (2021)
3. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: CVPR (2019)
4. Simon, C., Koniusz, P., Harandi, M.: On learning the geodesic path for incremental learning. In: CVPR (2021)

Table 1. Results on instance incremental split of CI-CC2M dataset at final step: The top-1 accuracy over various downstream datasets on zero-shot image classification task.

Methods	ImageNet	CIFAR-10	CIFAR-100	Caltech101	SUN397	Food101	Flowers102	DTD	Average
Joint	29.97	51.94	26.04	65.2	31.79	23.71	19.44	12.82	32.61
ER [1]	16.38	22.23	9.85	36.84	17.72	13.75	12.78	10.8	17.54
UCIR [3]	16.94	24.93	7.18	39.60	18.08	14.38	12.31	7.45	17.61
Co^2L [2]	17.15	22.24	8.23	45.08	17.9	14.44	11.41	8.83	18.16
GeoDL [4]	16.25	21.84	8.57	42.15	17.89	15.96	10.76	**10.64**	18.01
IncCLIP	**21.29**	**31.41**	**13.88**	**53.22**	**24.23**	**17.33**	**13.20**	10.20	**23.10**

Table 3. Image-text Retrieval Performance on instance incremental split of CI-CC2M dataset at final step: Zero-shot Image-Text Retrieval on MSCOCO and Flickr30k datasets with various methods. R@K means top-K recall.

Methods	Flickr30K						MSCOCO					
	image-to-text			text-to-image			image-to-text			text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Joint	35.7	62.4	71.8	25.16	50.52	62.36	18.8	41.82	52.94	13.14	30.74	41.78
ER [1]	15.9	39.7	51.5	12.3	29.32	38.32	10.36	25.34	35.94	6.59	18.23	26.22
GeoDL [4]	17.8	39.0	52.4	12.5	28.82	39.28	9.64	25.16	35.17	6.40	18.17	26.87
UCIR [3]	18.6	42.6	54.3	12.9	30.80	41.46	11.48	27.14	38.13	6.97	19.28	28.42
Co^2L [2]	18.5	41.7	53.2	12.8	30.12	40.22	10.08	26.4	36.04	6.69	18.78	27.50
IncCLIP	**21.4**	**48.9**	**62.1**	**16.68**	**38.72**	**50.28**	**13.48**	**31.22**	**41.08**	**8.76**	**23.30**	**33.11**

Table 4. Prompt Templates for image-text retrieval task.

Dataset	Task	Template
Flickr30K	image-to-text retrieval	a photo of {}
	text-to-image retrieval	a photo of {}
MSCOCO	image-to-text retrieval	a photo of {}
	text-to-image retrieval	a photo of {}