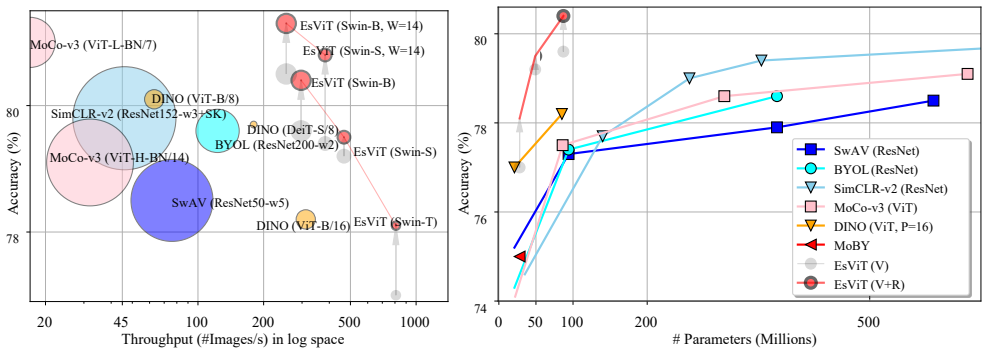

Figure 1: Efficiency vs accuracy comparison under the linear classification protocol on ImageNet. Left: Throughput of all SoTA SSL vision systems, circle sizes indicates model parameter counts; Right: performance over varied parameter counts for models with moderate (throughout/#parameters) ratio. EsViT pre-trained with and without the region-matching task are shown before and after the arrows, respectively. Please refer Section 4.1 for details.