

LIST OF FIGURES

2.1	Coarse-grained image classification vs fine-grained image classification	7
2.2	Vision transformer architecture	9
2.3	IELT architecture	11
2.4	Cross-layer refinement architecture	14
3.1	System model	16
3.2	Architecture of the first masking method	21
3.3	Architecture of the second masking method	22
4.1	Graphic accuracy and loss of combined method on CUB-200-2011 dataset	30
4.2	Graphic accuracy and loss of combined method on Stanford Dogs dataset	31
4.3	Visualization results from several modified method of IELT on all dataset	32