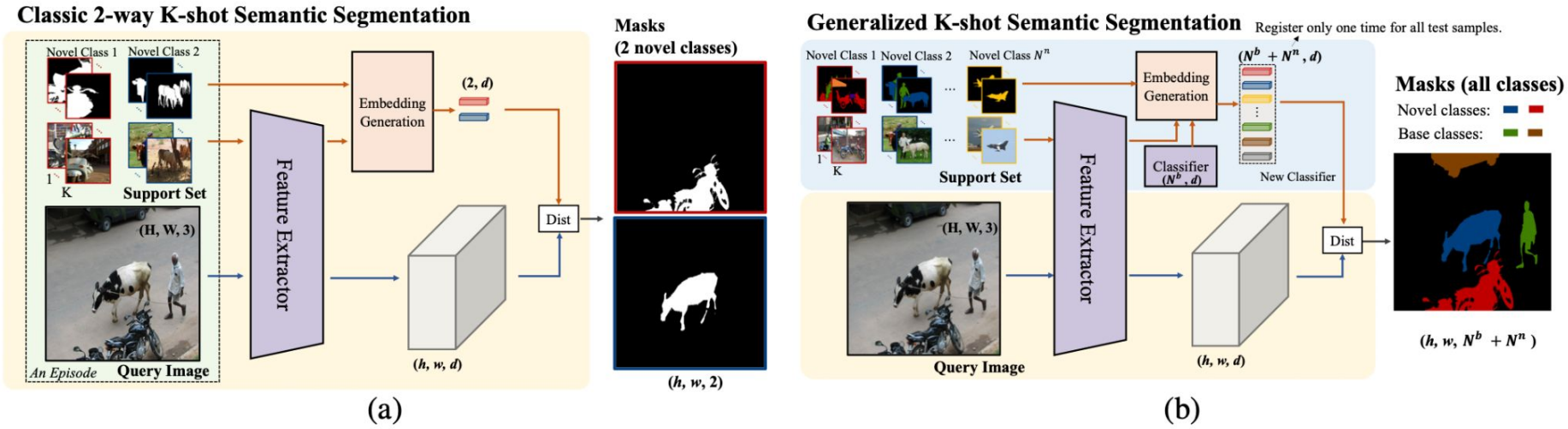
Applying ViT in Generalized Few-shot Semantic Segmentation

Liyuan Geng, Jinhong Xia, Yuanhe Guo, Instructed by Rob Fergus

Related Works



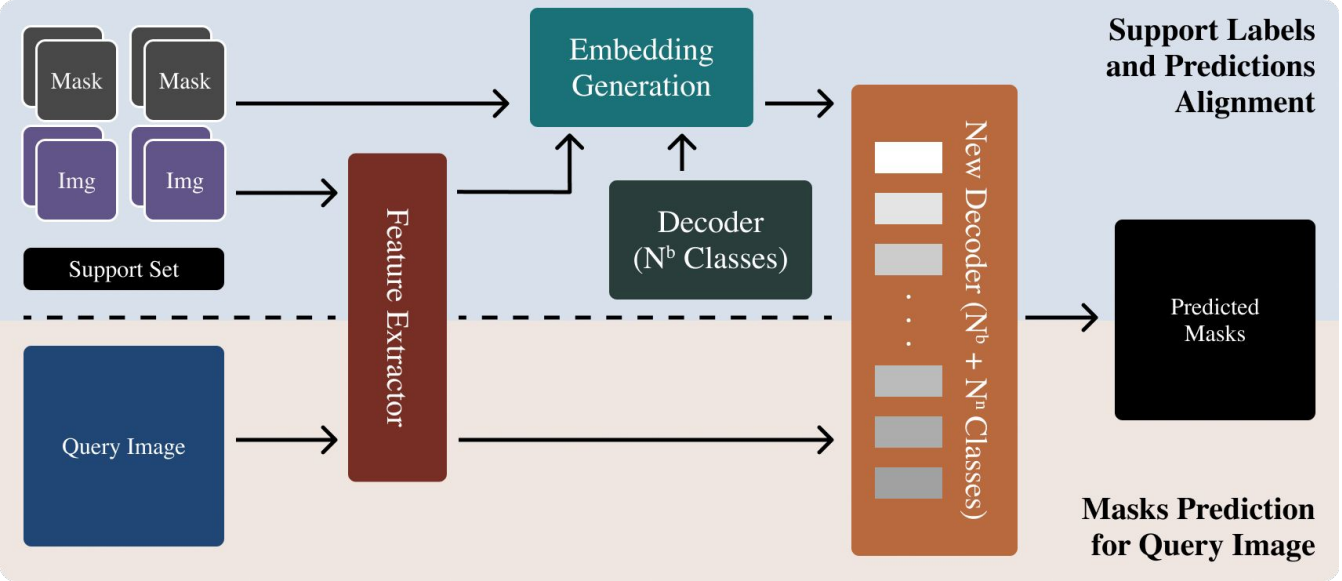
(Generalized Few-shot Semantic Segmentation, CVPR 2022)

Our Framework

Novelty:

1) Experimenting various combinations of feature extractors and classifiers

2) Applying pretrained ViT-based model in GFSS Framework



Leaderboard

Conclusion:

1) Dino V2 is a good few-shot learner

2) Mask Transformer as the decoder overfits easily in GFSS

		PASCAL						
Encoder	Decoder	Base Training	1-Shot			5-Shot		
			Base	Novel	Mean	Base	Novel	Mean
ResNet34 [13]	Linear	0.5539	0.2487	0.0846	0.1667	0.1546	0.0593	0.1070
ResNet34 [13]	UperNet [27]	0.6736	0.4650	0.0722	0.2686	0.4735	0.0635	0.2685
ResNet50 [13]	Linear	0.5939	0.3763	0.1260	0.2512	0.1705	0.0680	0.1193
ResNet50 [13]	UperNet [27]	0.7038	0.4739	0.0840	0.2790	0.4635	0.0726	0.2681
DINO [4]	Linear	0.5747	0.1240	0.1332	0.1286	0.1883	0.1777	0.1830
DINO [4]	Mask Transformer [7]	0.6803	0.6378	0.0806	0.3592	0.6437	0.2512	0.4475
DINOv2 [19]	Linear	0.7606	0.7246	**0.4799**	**0.6022**	0.6857	**0.5220**	0.6039
DINOv2 [19]	Mask Transformer [7]	**0.7987**	**0.8171**	0.2949	0.5560	**0.8217**	0.4598	**0.6407**