PolyphonicFormer : Unified Query Learning for Depth-aware Video Panoptic Segmentation

Haobo Yuan[1*], Xiangtai Li[2*], Yibo Yang[3], Guangliang Cheng[4], Jing Zhang[5], Yunhai Tong[2], Lefei Zhang[1], Dacheng Tao[3]

[1]Wuhan University, [2]Peking University, [3]JD Explore Academy, [4]SenseTime Research, [5]The University of Sydney.

Paper GitHub

1. Task Introduction

Depth-aware Video Panoptic Segmentation (DVPS):

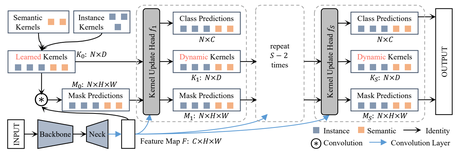
1). Taking raw videos as input.

2). Predicting instance-level temporal-consistent segmentation results.

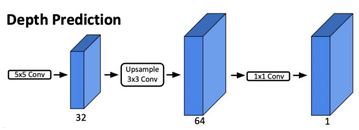
3). Predicting depth results for every pixel.

4). A complex and holistic scene understanding task.

2. Background and Motivation



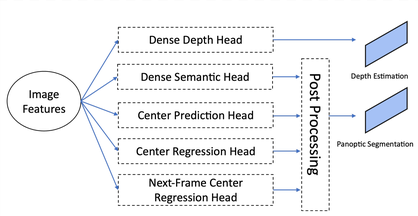
K-Net (NeurIPS '21)
Unified Query Learning for Segmentation



ViP-Deeplab (CVPR '21)
Dense Head for Depth Estimation

Though progress has been made in segmentation with query learning, the recent works on depth estimation still uses dense head like the first depth estimation method with deep learning in 2014.

Our target is to build a UNFIED framework to JOINTLY predict geometry and semantics for holistic scene understanding.



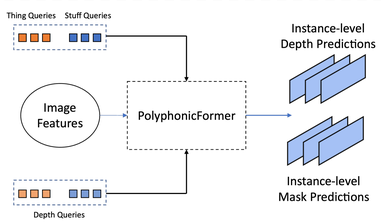
Previous Work on DVPS:

1). Complex.

2). Computationally heavy.

3). Ignore relationship between geometry and semantics.

4). Task competition.



PolyphonicFormer (Ours)

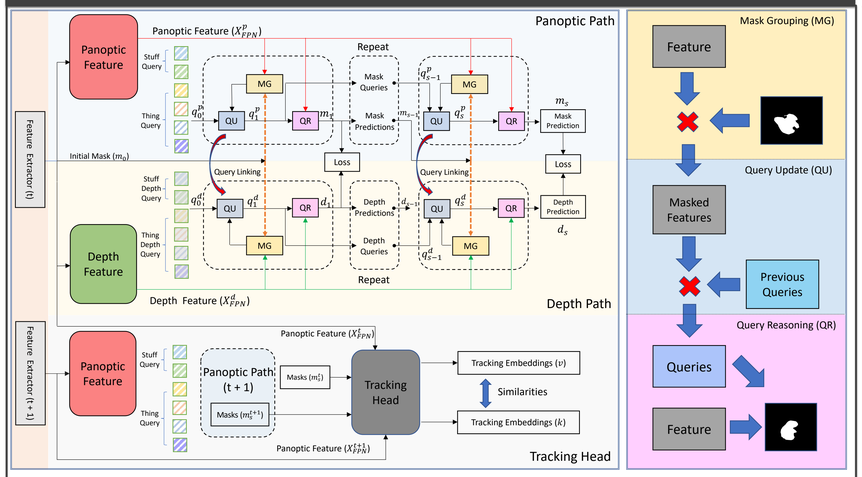
1). **Simple Pipeline**. ✅

2). **Unified and Efficient.** ✅

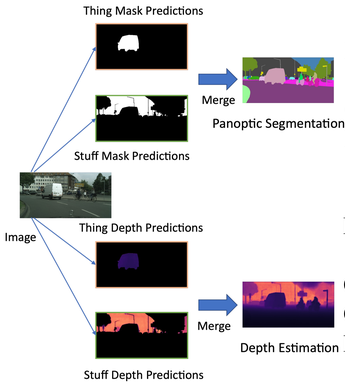
3). **Jointly predicting geometry and semantics.** ✅

4). **Mutual benefit.** ✅

3. Method



1). Joint depth and segmentation prediction.



Panoptic Segmentation: (with query learning)

One Query -> One Thing / Stuff
Bipartite Matching for calculating training loss.

Depth Estimation: (with query learning)

One Query -> Depth Map of Thing / Stuff
Cal Training Loss using Panoptic Bipartite Matching Results

Summary : Instance-level Depth Estimation Paradigm with Query Learning for implicitly leveraging semantic information.

2). Unified framework for mutual benefit.

We adopt query learning between the two paths for mutual benefit.

3). Query Modeling for encoding object information.

We iteratively update the object queries for both panoptic and depth paths to extract instance-level information.

4). Tracking Head for temporal alignment.

We perform tracking beyond the learned object queries. $\mathcal{L}_{\text{track}} = \log[1 + \sum_{\boldsymbol{k}^+} \sum_{\boldsymbol{k}^-} \exp(\boldsymbol{v} \cdot \boldsymbol{k}^- - \boldsymbol{v} \cdot \boldsymbol{k}^+)]$.

4. Experiments

DVPQ$^k_\lambda$ on Cityscapes-DVPS	k = 1		k = 2		k = 3		k = 4		Average		FLOPs
PolyphonicFormer $\lambda = 0.50$	70.6	63.0	76.0	62.9	49.2	72.9	59.3	42.3	71.7	56.5	36.9	70.8	62.3	47.9	72.9	-
PolyphonicFormer $\lambda = 0.25$	67.8	61.0	72.8	60.4	47.6	69.8	56.9	40.8	68.6	54.3	35.8	67.8	59.9	46.3	69.8	-
PolyphonicFormer $\lambda = 0.10$	50.2	43.4	55.2	44.4	33.4	52.4	41.5	28.6	51.0	39.5	24.7	50.4	43.9	32.5	52.3	-
Average: PolyphonicFormer	62.9	55.8	68.0	55.9	43.4	65.0	52.6	37.2	63.8	50.1	32.5	63.0	**55.4**	42.2	65.0	1,675G
Average: ViP-Deeplab [43]	61.9	55.9	66.3	55.6	44.3	63.8	52.4	38.4	62.6	50.4	34.6	61.9	55.1	43.3	63.6	9,451G

DVPQ$^k_\lambda$ on SemKITTI-DVPS	k = 1		k = 5		k = 10		k = 20		Average		FLOPs
PolyphonicFormer $\lambda = 0.50$	58.5	55.1	61.0	52.0	42.3	59.1	50.6	39.9	58.5	49.9	38.6	58.0	52.8	44.0	59.2	-
PolyphonicFormer $\lambda = 0.25$	56.3	54.0	57.9	49.7	41.1	56.0	48.4	38.7	55.5	47.7	37.6	55.0	50.5	42.9	56.1	-
PolyphonicFormer $\lambda = 0.10$	41.8	41.1	42.4	35.1	28.2	40.1	33.7	26.0	39.3	33.0	25.1	38.7	35.9	30.1	40.1	-
Average: PolyphonicFormer	52.2	50.1	53.8	45.6	37.2	51.7	44.2	34.9	51.1	43.4	33.8	50.6	**46.4**	39.0	51.8	402G
Average: ViP-Deeplab [43]	48.9	42.0	53.9	45.8	36.9	52.3	44.4	34.6	51.6	43.4	33.0	51.1	45.6	36.6	52.2	2,267G

Results on Cityscapes-DVPS and SemKITTI-DVPS (DVPQ).
Our method achieves better results with about ¼ computational cost.

Method	k = 1	k = 2	k = 3	k = 4	VPQ
VPSNet [21]	65.0	57.6	54.4	52.8	57.5
SiamTrack [63]	64.6	57.6	54.2	52.7	57.3
ViP-Deeplab [43]	69.2	62.3	59.2	57.0	61.9
Ours (ResNet50)	65.4	58.6	55.4	53.3	58.2
Ours (Swin-b)	70.8	63.1	59.5	56.8	62.3

Method	*DSTQ*
rl-lab	54.8
yang26	55.6
Vip-Deeplab	63.3
PolyphonicFormer	63.6
PolyphonicFormer*	64.6

Results on Cityscapes-VPS. (VPQ)
Our method also outperforms some other works on VPS.

Method	Depth	Panoptic	Ins	PQ ↑	abs rel ↓
ViP-Deeplab [43]	✓	✓	-	60.6	0.112
Depth	✓	-	-	N/A	0.084
Panoptic	-	✓	-	63.7	N/A
Hybrid (ours)	✓	✓	-	65.1	0.089
PolyphonicFormer (ours)	✓	✓	✓	65.2	0.080

ICCV-2021 SemKITTI-DVPS Challenge. PolyphonicFormer is the WINNER.

L_{depth}	PQ ↑	abs rel ↓
0.1	65.4	0.101
1.0	65.3	0.089
5.0	65.2	0.080
10	65.4	0.079

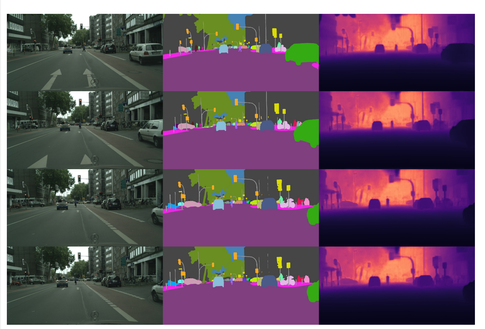
Unified framework is good for mutual benefit and robust to loss weight choices between sub-tasks rather than mutual competition.

Stages	PQ ↑	abs rel ↓
1	64.1	0.081
2	64.6	0.081
3	65.2	0.080

Method	DSTQ ↑	AQ ↑
PolyphonicFormer + DeepSort [62]	51.8	25.9
PolyphonicFormer + Unitrack [59]	49.3	22.5
PolyphonicFormer + QuasiDense [38]	63.6	46.2

Iteratively query updating instance-level information.

PolyphonicFormer is capable of tracking with different appearance-based tracking heads.



The output of the Depth-aware Video Panoptic Segmentation with PolyphonicFormer.

Instance-level Depth Estimation for leveraging semantic information.

ECCV TEL AVIV 2022