Balancing Shared and Task-Specific Representations: A Hybrid Approach to Depth-Aware Video Panoptic Segmentation

Kurt H.W. Stolle

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

BACKGROUND

Depth-aware video panoptic segmentation (DVPS) combines segmentation, depth estimation, and object tracking in video. Such information has a critical role in autonomous driving and robotics applications.

Current approaches share information across tasks either explicitly, *i.e.* modeling interactions between task-specific embeddings [4, 5], or implicitly through a shared object representation [2].

We propose **Multiformer**, a hybrid architecture that combines task-specific and shared object representations ('queries'). Furthermore, we show that this architecture can be extended with metric depth estimation. Finally, a design space exploration on various query decoder designs is provided.

KEY COMPONENTS

- **Mask transformer model** with masked-attention [1] extended to depth estimation and tracking.
- **Branched decoder block** with task-specific query refinement branches.
- **Context adapter** that seeds the initial (shared) queries.
- **Depth head** that directly estimates metric depth without min-max denormalization.

DATASETS & TRAINING

Name	Train videos (frames)	Val. videos (frames)	Annotations
Cityscapes-DVPS [4]	400 (12,000)	50 (1,500)	Every 5th frame
SemKITTI-DVPS [4]	10 (19,130)	1 (4,071)	Sparse reprojected

Models were trained on **4 NVIDIA H100 GPUs** for **20K steps** in **batches of 32 annotated images**. Furthermore, the **AdamW optimizer** was adopted, having **0.0005 peak learning rate** in a polynomial decay schedule. Additionally, random color jitter and horizontal flip augmentations were applied.

MAIN RESULTS

Method	Backbone	DVPQ ↑ [%]			Depth error ↓	
		All	Thing	Stuff	Abs.Rel.	RMSE
ViP-DeepLab [4]	ResNet-50	42.0	27.6	51.5	0.070	3.67
MonoDVPS [3]	ResNet-50	48.8	31.0	61.7	0.070	3.67
PolyphonicFormer [5]	ResNet-50	48.1	35.6	57.1	0.081	4.01
UniDVPS [2]	ResNet-50	51.8	37.1	62.5	0.067	3.88
Multiformer (ours)	ResNet-50	**54.8**	**37.4**	**67.4**	**0.066**	**3.35**
PolyphonicFormer [5]	Swin-B	55.4	43.3	63.6	0.065	3.80
Multiformer (ours)	Swin-B	**59.4**	**46.0**	**69.2**	**0.048**	**2.81**

State-of-the-art performance on *Cityscapes-DVPS* [4], surpassing previous methods by **3.0** (ResNet-50) and **4.0** (Swin-B) depth-aware video panoptic quality (DVPQ) %-points and improving depth estimation accuracy. Measured using Multiformer with $N_B = 9$ decoder blocks (large).

CONCLUSION & IMPACT

- **Hybrid architecture.** Introduces a hybrid decoder that effectively balances shared and task-specific representations through a *context adapter* that produces initial queries from task-specific image features and using a *branched decoder block* that enables task-specific decoding while maintaining a shared representation between blocks.
- **Metric depth estimation.** Incorporates an innovative depth estimation head that eliminates the dataset-specific depth hyper-parameters and improves accuracy.
- **Design insights.** Provides valuable insights into multi-task decoder architectures, demonstrating the benefits of task-specific branching with shared interfaces for complex vision tasks.

REFERENCES

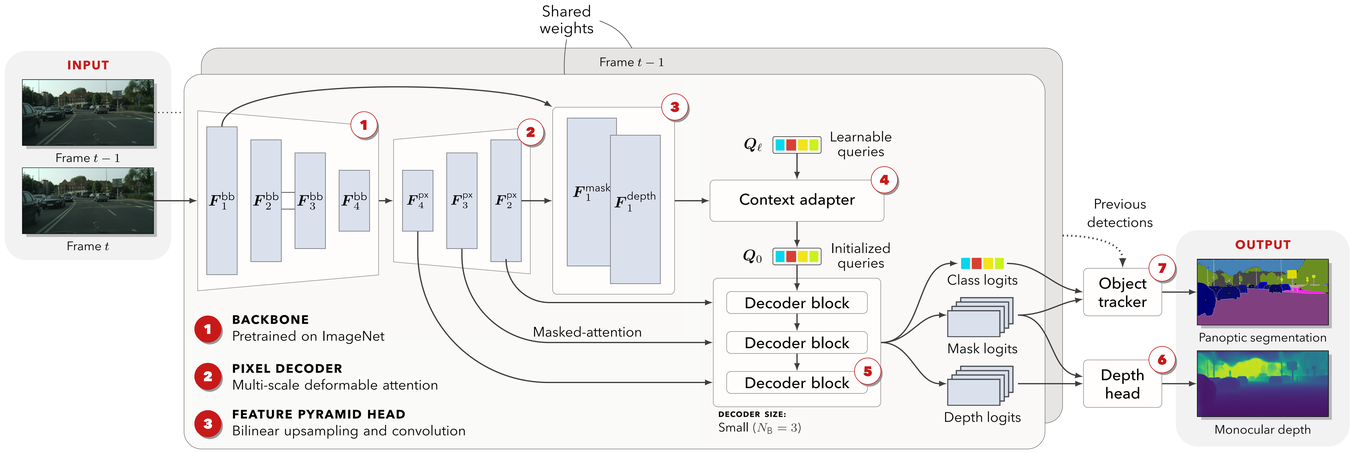
[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022.

[2] Kim Ji-Yeon, Oh Hyun-Bin, Kwon Byung-Ki, Dahun Kim, Yongjin Kwon, and Tae-Hyun Oh. UniDVPS: Unified model for depth-aware video panoptic segmentation. *IEEE Robotics and Automation Letters*, pages 1–8, 2024.

[3] Andra Petrovai and Sergiu Nedevschi. MonoDVPS: A self-supervised monocular depth estimation approach to depth-aware video panoptic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3077–3086, 2023.

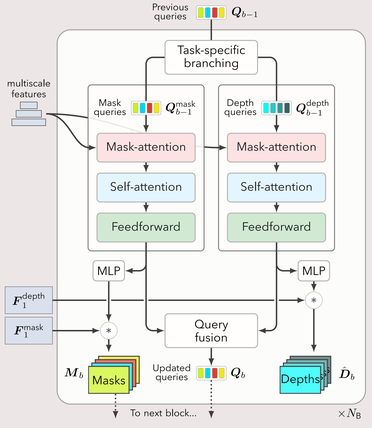
[4] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. ViP-DeepLab: Learning visual perception with depth-aware video panoptic segmentation. In *CVPR*, 2020.

[5] Haobo Yuan, Xiangtai Li, Yibo Yang, Guangliang Cheng, Jing Zhang, Yunhai Tong, Lefei Zhang, and Dacheng Tao. PolyphonicFormer: Unified query learning for depth-aware video panoptic segmentation. In *ECCV*, 2022.



BRANCHED DECODER BLOCK

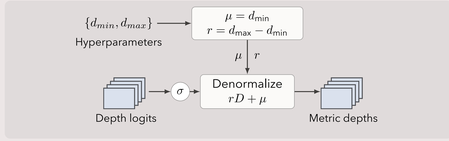
5



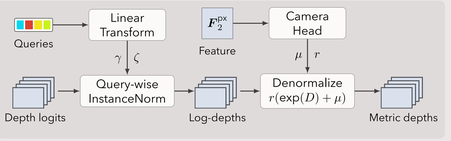
1. **Learnable queries are split** into task-specific branches through a learnable linear transform.
2. **Each branch incorporates task-specific nuances** via masked-attention, self-attention and a feed-forward layer.
3. **Task-specific queries are fused** at the interface between blocks via a linear transform and normalization.

MONOCULAR DEPTH ESTIMATION

6



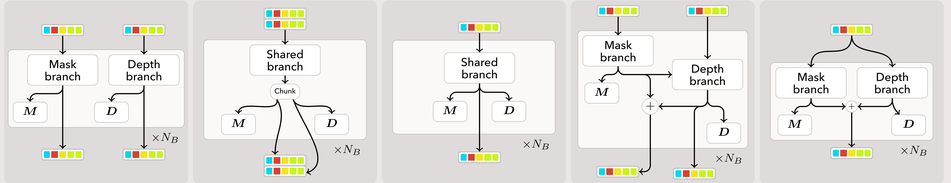
Min-max denormalization head [4, 5, 2]	Abs.Rel ↓	RMSE ↓
Multiformer (large)	0.069	3.54
– Weighted merge	0.073	3.74
– Context adapter	0.074	3.84



Direct metric depth head (ours)	Abs.Rel ↓	RMSE ↓
Multiformer (large)	0.066	3.35
– Weighted merge	0.076	3.81
– Context adapter	0.078	3.86

DECODER BLOCK DESIGN SPACE

Exploration of various decoder block designs on the overall depth-aware video panoptic quality (DVPQ) on *Cityscapes-DVPS* [4] with $N_B = 3$ blocks.



Separated design
DVPQ $= 45.7$ ($N_P = 17.8$ M)

Concatenated design
DVPQ $= 51.2$ ($N_P = 19.8$ M)

Unified design
DVPQ $= 52.3$ ($N_P = 17.6$ M)

Sequential design
DVPQ $= 52.4$ ($N_P = 18.2$ M)

Hybrid design (ours)
DVPQ $= 52.7$ ($N_P = 18.0$ M)

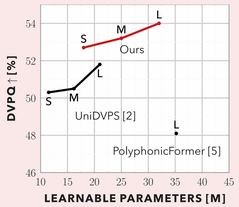
MODEL SCALING

The proposed model is scaled through the amount of decoder blocks N_B.

Small
3 blocks

Medium
6 blocks

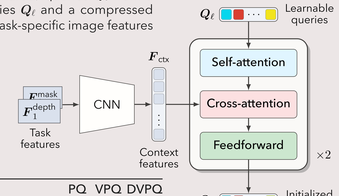
Large
9 blocks



CONTEXT ADAPTER

4

Produces the initial (shared) queries Q_0 from a set of learnable queries Q_ℓ and a compressed representation of the task-specific image features $\{F_1^{mask}, F_1^{depth}\}$.

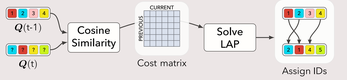


Ablation	PQ	VPQ	DVPQ
Passthrough ($Q_0 = Q_\ell$)	65.1	57.1	52.3
Context adapter	65.2	57.5	52.7

Evaluated on Cityscapes-DVPS with $N_B = 3$ (small).

OBJECT TRACKER

7



Method	Backbone	DVPQ$^\lambda_\kappa$		κ=1	κ=5	κ=10	κ=20	Avg.
ViP-DeepLab	WR-50			48.9	45.8	44.4	43.4	45.6
PolyphonicFormer [5]	Swin-B		λ=0.50	58.5	52.0	50.6	49.9	52.8
			λ=0.25	56.3	49.7	48.4	47.7	50.5
			λ=0.10	41.8	35.1	33.7	33.0	35.9
			Avg.	52.2	45.6	44.2	43.4	46.4
Multiformer (ours)	Swin-B		λ=0.50	56.6	55.2	54.6	49.6	54.0
			λ=0.25	51.4	49.6	49.5	48.5	49.7
			λ=0.10	49.1	47.2	46.6	46.2	47.3
			Avg.	**52.3**	**50.6**	**50.2**	**48.2**	**50.3**

Depth-aware video panoptic quality on *SemKITTI-DVPS* [4] for varying window size (κ) and depth threshold (λ) using $N_B = 9$ decoder blocks (large). The proposed method achieves state-of-the-art performance, and exhibits less DVPQ degradation at larger window sizes than previous methods.

CODE & MODELS

research.khws.io/multiformer

CONTACT

Kurt H.W. Stolle
kurt@computer.org

This publication is part of the NEON project with file number 17628 of the Crossover research program, which is (partly) financed by the Dutch Research Council (NWO). The Dutch national compute infrastructure was used with the support of the SURF Cooperative using grant EINF-5438.