Unsupervised depth prediction from monocular sequences: Improving performances through instance segmentation

Ambroise Moreau, Matei Mancas, Thierry Dutoit – name.surname@umons.ac.be
Isia Lab, University Of Mons, Belgium

Architecture details:

- Our model uses soft-attention modules (ae and ad) to extract task-specific features from a set of shared features as in [1].
- The shared encoder is a pretrained Resnet18 while the shared decoder is the depth decoder of monodepth2 [2].
- For the pose, we use the pose decoder of monodepth2 without attention decoders.

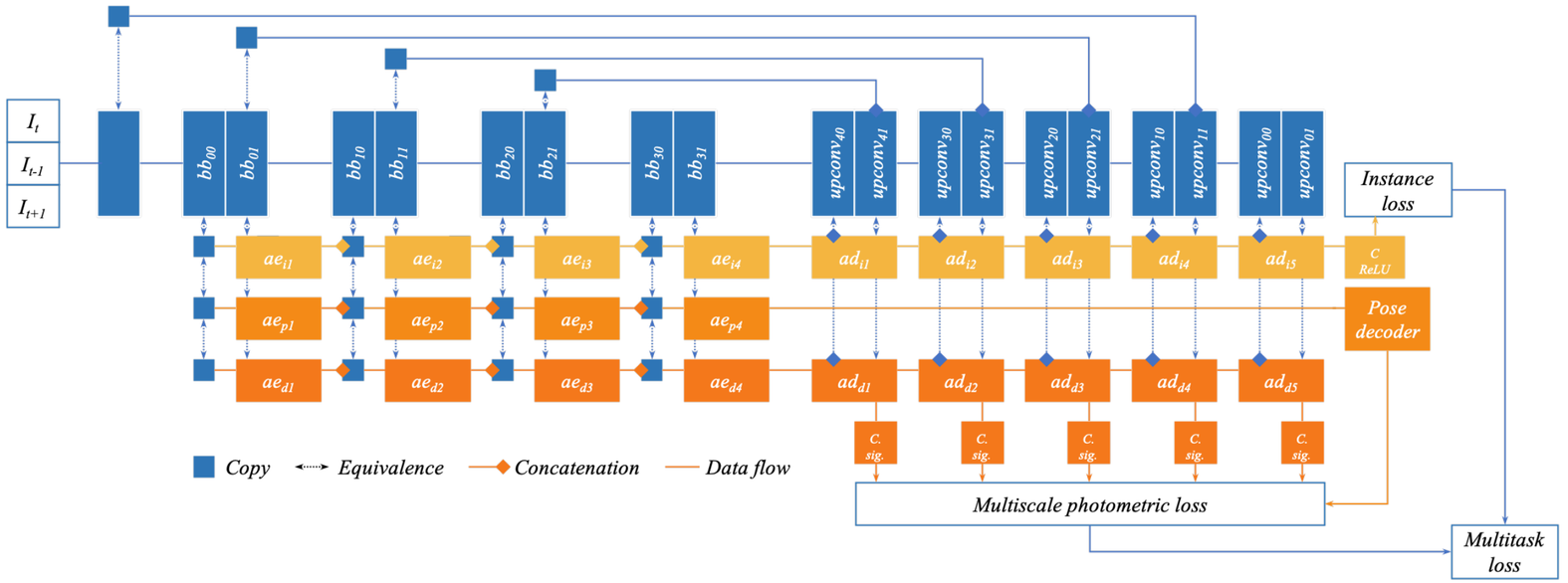
Training details:

- Our model is trained on Kitti and Cityscapes. The training sets contain 39810 triplets and the validation sets count 4424 triplets.
- The ground truth for the instance segmentation task is created with Mask R-CNN trained on Cityscapes.
- The number of epochs is set to 20.
- Model trained in single task (ST) and multi-task (MT) mode. Depth prediction and pose estimation are complementary and are always solved together.

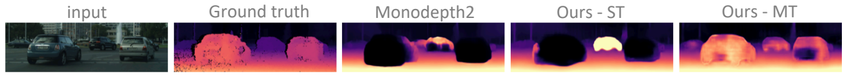
Losses:

- The depth prediction and pose estimation are optimised thanks to the multiscale photometric loss presented in [2].
- Instance segmentation is achieved by computing pixels embeddings (i.e. x and y offset from the pixels to the object's centroid coordinates). These embeddings need to be clustered with an algorithm such as OPTICS [3].
- Instance segmentation is supervised by the L2-norm between the prediction and the ground truth.
- The two losses are combined into a multi-task loss which implements equal weighting, Dynamic Weight Average [1] or homoscedastic weighting [4].

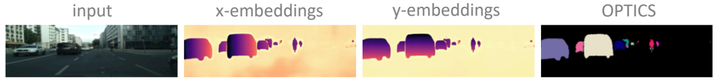
Our model solves unsupervised depth prediction, pose estimation and instance segmentation at once.



Results:



Our experiments show that combining the three tasks improves depth estimation in challenging situations such as objects moving at the same speed as the camera.



The embeddings computed by the model can be used with OPTICS to obtain the final instances.

References:

[1] Liu, Shikun, Edward Johns, and Andrew J. Davison. "End-to-end multi-task learning with attention." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[2] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." Proceedings of the IEEE International Conference on Computer Vision. 2019.
[3] Ankerst, Mihael, et al. "OPTICS: ordering points to identify the clustering structure." ACM Sigmod record 28.2 (1999): 49-60.
[4] Kendall, Alex, Yarin Gal, and Roberto Cipolla. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.