

Supplementary Material for MILA: Multi-Task Learning from Videos via Efficient Inter-Frame Attention

Donghyun Kim¹, Tian Lan², Chuhan Zou², Ning Xu²,
Bryan A. Plummer¹, Stan Sclaroff¹, Jayan Eledath², Gerard Medioni²

¹Boston University, ² Amazon

¹{donhk, bplum, sclaroff}@bu.edu, ²{tianlan, ninxu, zouchuha, eledathj, medioni}@amazon.com

A. Experiment Details for Fair Comparison

As stated in Sec.4.2 in the main paper, for fair comparison to other methods (MTAN [9], Cross-Stitch [10] and MultiNet++ [2]), we unify the backbone for all method as Deeplab-ResNet101 [1, 5]. Note that after using Deeplab-ResNet101, the performance for MTAN, Cross-Stitch and MultiNet++ gets improved compared to their originally reported numbers as shown in Table A. Deeplab-ResNet101 is initialized with pre-trained weights on ImageNet but decoders use randomly initialized weights. Then, the model is finetuned for multi-task learning.

For implementation details, since all multi-task learning models except Cross-Stitch use shared encoder, we pre-train a shared multi-task encoder based on Deeplab-ResNet101 for each method (named as “D101-MultiTask”). For the Cross-Stitch model we pre-train a task specific encoder (named as “D101-SingleTask”). Using networks pre-trained on multi-task learning shows better performances than just directly finetuning pre-trained weights on ImageNet [4]. We first train D101-SingleTask and D101-MultiTask with Adam optimizer with learning [8] rate $1e^{-4}$ with a decay rate $1e^{-1}$. For our SlowFast based network, we use the D101-MultiTask for our slow network. We also train a D18-MultiTask for the *Fast* network with the same optimizer and training setting. For fair comparison with MultiNet++, we use the same pre-trained slow-fast network for MultiNet++, but concatenate neighboring frames as input to learn from videos as proposed in their paper.

Note that we also report in the main paper the performance comparison with our pre-trained “D101-SingleTask” and “D101-MultiTask”.

A.1. Weighting Strategies.

In the main paper, we use the equal weighting for each task. In Table B, we show performance with different weighting strategies for the multiple tasks: Uncertainty Weighting [7] and Dynamic Weight Average [9]. These weighting strategies are proposed to find a balance between

Model	Segmentation		Depth	
	mIOU \uparrow	Acc. \uparrow	Abs. \downarrow	Rel. \downarrow
MTAN-SegNet [9]	53.0	91.1	1.44	33.6
MTAN-ResNet101	64.2	94.5	1.06	26.3
Cross-Stitch-SegNet [9]	50.1	90.3	1.54	34.5
Cross-Stitch-ResNet101	64.5	94.5	1.04	33.0

Table A: Comparison of different backbones. Using the ResNet101 backbone improves the accuracy by a large margin.

Backbone	Weighting	mIOU (\uparrow)	Depth Err. (\downarrow)
D101-18	Equal	64.3	1.02
D101-18	Uncertainty [7]	64.3	1.01
D101-18	DWA [9]	64.2	1.02

Table B: Different weighting strategies for MILA. MILA is not sensitive to the weighting strategies.

different tasks, since a model can be biased to a certain task. A desired multi-task learning model should not depend on these weighting schemes, so that the model itself can find a proper balance between tasks. We observe that MILA achieves the similar performance on different weighting strategies and thus is not sensitive to the weighting schemes.

B. Ablation for *Slow* network-only

We apply our full feature propagation method (inter-frame local attention (ILA) + multi-frame feature propagation + task specific) on *Slow* network-only model. This means that the keyframe interval is 1. Table C shows the results on Cityscapes. In all cases, using our feature propagation method shows significant improvements on the depth estimation task.

Model	Segmentation		Depth	
	mIOU \uparrow	Acc. \uparrow	Abs. \downarrow	Rel. \downarrow
D101-Multi	63.8	94.4	1.06	31.9
D101-Multi + Ours	64.4	94.6	1.02	25.8
D50-Multi	63.1	94.2	1.09	33.0
D50-Multi + Ours	63.3	94.5	1.02	25.1
D18-Multi	60.3	93.4	1.21	34.8
D18-Multi + Ours	61.5	93.8	1.07	25.8

Table C: Evaluation of a Slow-only network with our task-specific inter-frame local attention (ILA) module on Cityscapes.

C. Additional Comparison for Feature Propagation

In Table D, we provide more detailed comparisons between our attention based feature propagation and the optical flow based feature warping method by Jain *et al.* [6]. We compare the task of video semantic segmentation as in [6]. Our method outperforms [6], and is more robust to different keyframe interval. Note that we only apply our ILA module for fair comparison with other feature propagation methods (*i.e.* without multi-frame and task-specific attention).

Backbone	K	Feature Prop.	mIOU (%)
D101-50	5	Optical flow	74.2
D101-50	5	ILA (Ours)	75.1
D101-50	10	Optical flow	72.9
D101-50	10	ILA (Ours)	74.8

Table D: Supplements to Table 5 in the main paper. Comparison with optical-flow based feature propagation [6] for the semantic segmentation task on Cityscapes. A keyframe interval is denoted by K . We show detailed comparison using different backbones and keyframe intervals.

D. Additional Qualitative Results

In the main paper, we only put the qualitative results on Cityscapes due to the limited space. We also provide the qualitative results on NYU v2 in Figure 1. Compared to MultNet++ [2], our method is robust to non-keyframes and obtains similar performances as the computationally heavy method [9]. In addition, we also provide the supplemental video of predictions of the *Slow* and *Fast* networks, where the *Fast* network produces qualitatively similar performances as the *Slow* network.

E. Discriminator Details

The discriminator D takes feature maps from the last residual block of slow and fast network. The discrimina-

tor consists of three 3x3 convolutional layers with relu activation, a global average pooling layer, and sigmoid activation, which outputs a single value. It should be noted that the discriminator is only used during training, therefore it does not increase GFLOPs in inference time. [R4] We update the encoder and discriminator jointly by using a gradient-reversal layer [3] for Eq. 3.

References

- [1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1
- [2] Sumanth Chennupati, Ganesh Sistu, Senthil Yogamani, and Samir A Rawashdeh. Multinet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2
- [3] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 2
- [4] Yuan Gao, Jiayi Ma, Mingbo Zhao, Wei Liu, and Alan L Yuille. Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3205–3214, 2019. 1
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1
- [6] Samvit Jain, Xin Wang, and Joseph E Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8866–8875, 2019. 2
- [7] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018. 1
- [8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1
- [9] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1871–1880, 2019. 1, 2
- [10] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016. 1

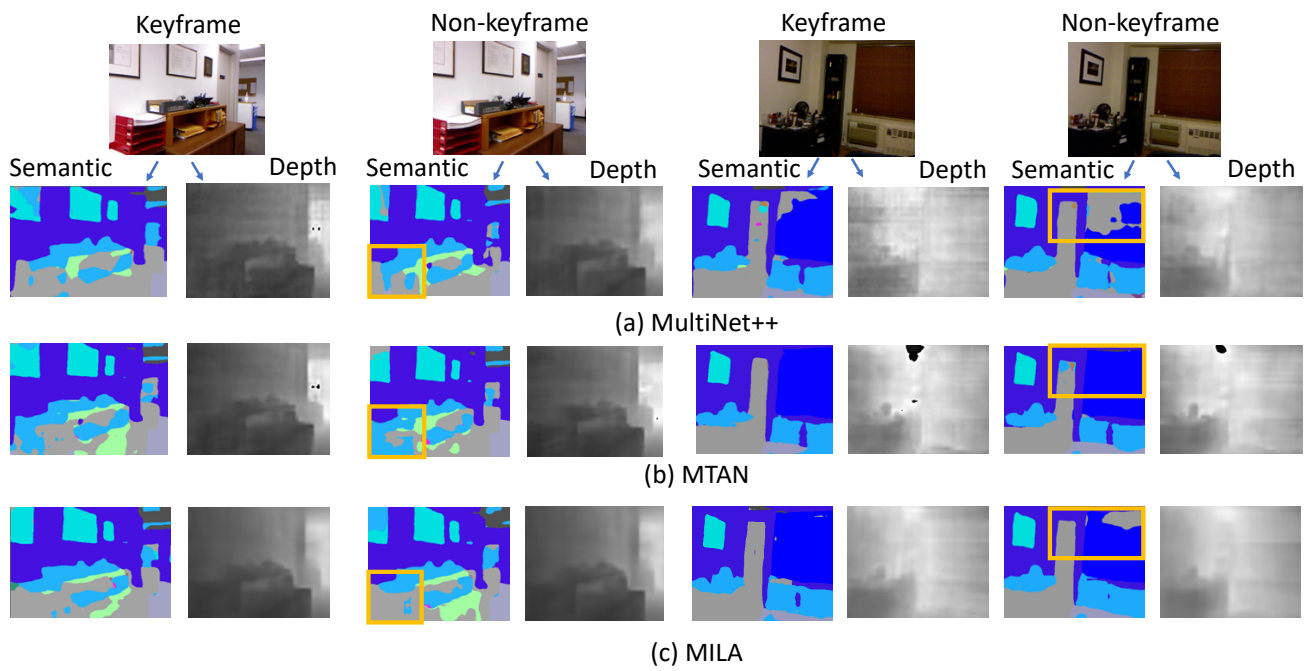


Figure 1: Qualitative results for multi-task learning on videos on the NYUd v2 dataset. We show comparison between (a) MultiNet++, (b) MTAN and (c) ours. MultiNet++ performs worse on the non-keyframe.