MBMT-Net: A Multi-Backbone, Multi-Task Deep Convolutional Neural Network

George Ciubotariu
Babeș-Bolyai University

WeADL 2022 Workshop

The workshop is organized under the umbrella of WeaMyL, project funded by the EEA and Norway Grants under the number RO-NO-2019-0133. Contract: No 26/2020.

Working together for a green, competitive and inclusive Europe

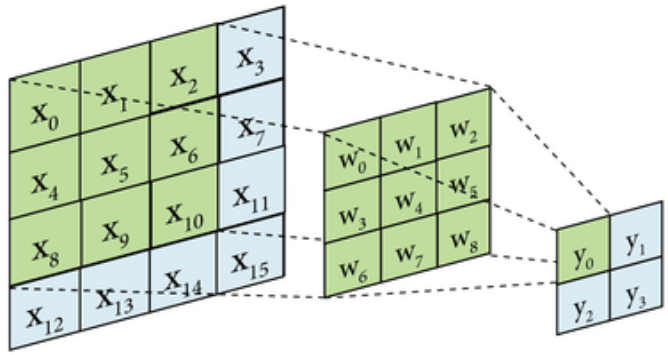
Contents

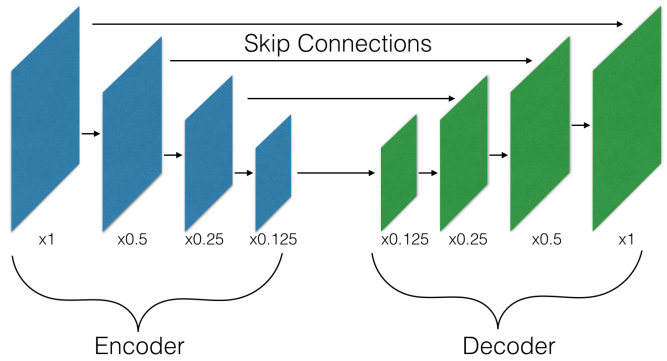
Problem statement

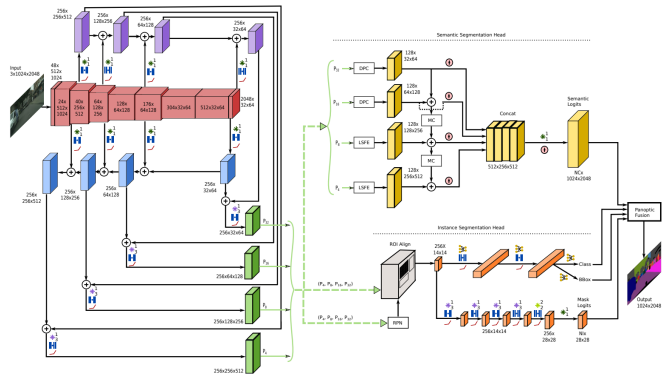
- Convolutional neural networks' architectures keep changing for the better
- However, substantial improvements could be achieved not only by using better structural components, but empowering the feature extract means
- Therefore, single-task learning contexts may be suboptimal for state-of-the-art neural networks, which determined us to look into this matter
- The relevance of learned patterns is critical in obtaining robust predictions, since by offering the model more diverse and specialised features, performance may increase without the need of architectural changes, apart from the addition of multiple task-specific decoders

Brief Background

- The Convolution
- Encoder-Decoder architecture
- Convolutional Neural Networks (CNN)







Research Questions and Original Contributions

- **RQ1**: *How to enhance the performance of dense tasks by using the multitask learning paradigm? In this respect, we are introducing the MBMT-Net model.*
  - Supervised learning based analysis on NYUv2 dataset [1, 2] for three dense tasks:
    - Semantic Segmentation (SS)
    - Depth Estimation (DE)
    - Surface Normals Prediction (SNP)
- **RQ2**: *To what extent does MBMT-Net improve the performance of current state-of-the-art approaches in dense tasks?*
  - Comparison with multiple model variants from the literature
    - Efficient-PS [3]
    - MT-Efficient-PS [3]
    - MTAN [1]

Related Work

- EfficientNet [4]
  - variants have been pretrained on ImageNet [5]
  - outputs multi-scale features
- EfficientPS [3]
  - employs a two-way Feature Pyramidal Network (FPN) [6]
  - outputs dense prediction cells [7] and residual pyramids
- CBNet [8, 9]
  - technique of assembling several identical backbones
  - include higher-level features into succeeding backbones
- MTAN [1]
  - state-of-the-art performance for multiple dense tasks
  - synchronises the tasks' convergence speed using a procedure named *Dynamic Weight Average*

Deep Learning Architectures

- [4] **Single-Task Models** (ST-Net)
  - low number of parameters
  - specialises in extracting features for the unique task
- [8, 9] **Composite-Backbone Models** (CB-Net)
  - boosts context understanding & receptive field via skip connections
  - constrained to use identical backbones, as the feature maps sharing cannot be performed otherwise
- [1] **Multi-Task Models** (MT-Net)
  - built using a hard or soft parameter sharing strategy
  - enhances the robust feature extraction abilities of a model
- **Multi-Backbone Multi-Task Models** (**Ours** - MBMT-Net)
  - combines the multiple benefits of each type of architecture
  - reduces number of parameters necessary to achieve good result
  - more robust pattern extraction means thanks to richer feature maps

NYU Depth Dataset V2

The NYU-Depth V2 data set is comprised of video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect. It features:

- 1449 densely labeled pairs of aligned RGB and depth images
- 464 new scenes taken from 3 cities
- Each object is labeled with a class and an instance number
- 288 x 384 pixels images
- pre-processed dataset (numpy arrays) offered by [1]

We are only using the densely labeled images provided by [1] in a pre-processed format. We have used the PyTorch [10] framework for implementing the model architecture.

Indoor Segmentation and Support Inference from RGBD Image

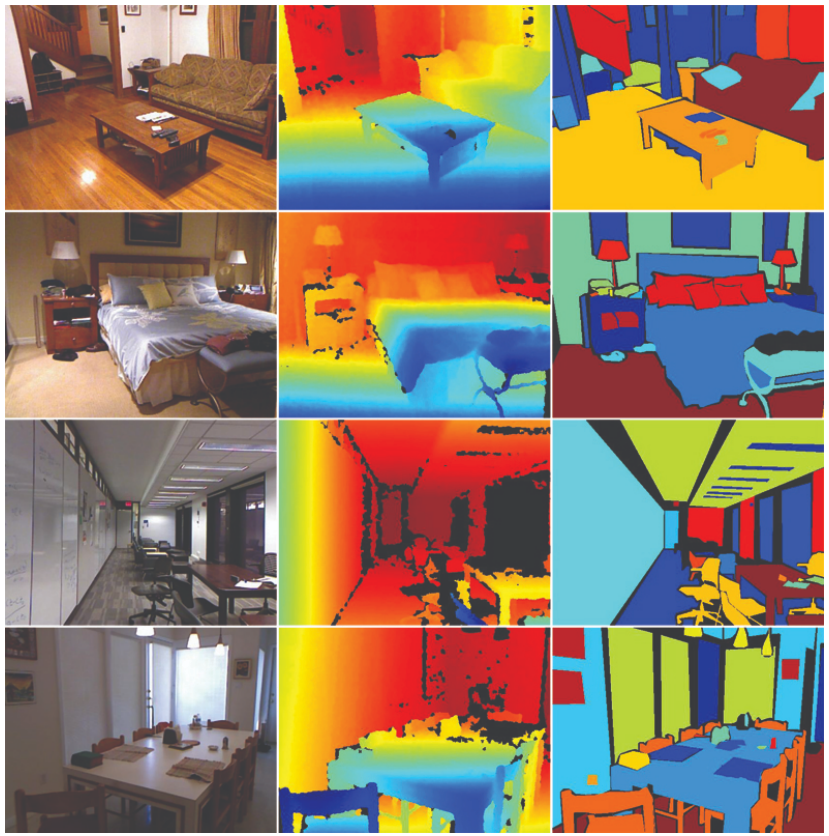
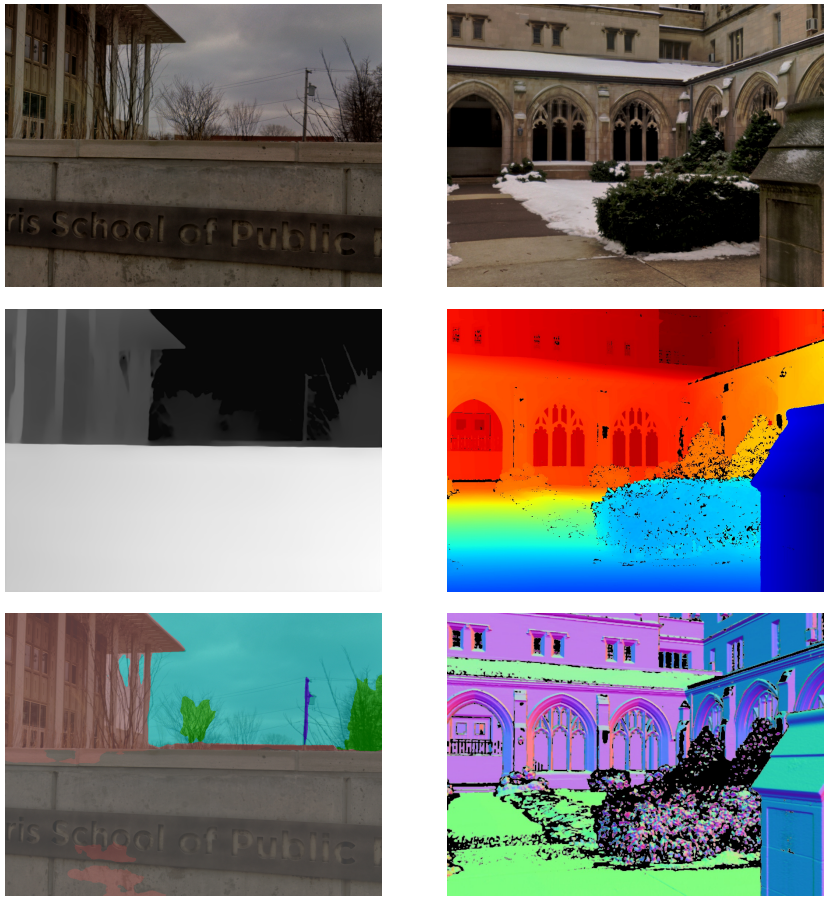


Figure: Sample images from NYUv2 dataset [2]

Deep Learning Tasks

- Semantic Segmentation
- Depth Estimation
- Surface Normals Prediction

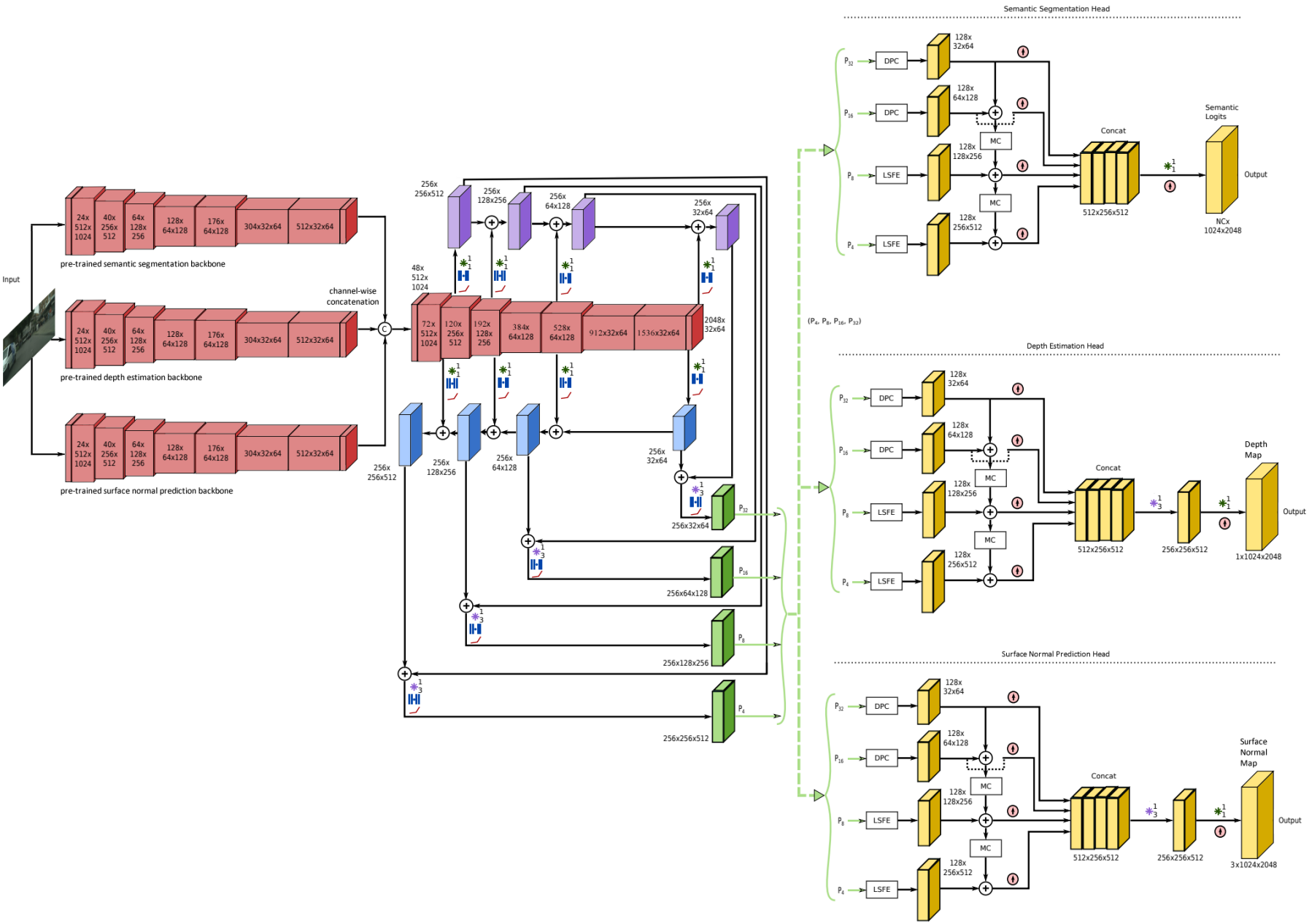


Models' Abbreviations

Context	**Backbone**	**Architecture and Tasks**	**Abbreviation**
Single-Task	1xEfficientNet-b2	EfficientPS (SS)	ST-SS-Net
Single-Task	1xEfficientNet-b2	EfficientPS (DE)	ST-DE-Net
Single-Task	1xEfficientNet-b2	EfficientPS (SNP)	ST-SNP-Net
Multi-Task	1xEfficientNet-b5	MT-EfficientPS DWA (SS+DE+SNP)	MT-Net
Multi-Backbone Multi-Task	3xEfficientNet-b2	MT-EfficientPS DWA (SS+DE+SNP)	MBMT-Net

Table: The architectures used in our experiments

Architectural Design



Training Methodology

- ST-Nets training (for SS/DE/SNP - using EfficientNet-b2 backbone)
- MT-Net training (with the larger EfficientNet-b5)
- full MBMT-Net training (with both EfficientNet-b2 and -b5)
  - with pre-trained encoders for each task
  - with encoders being trained from scratch
- partial MBMT-Net training (only EfficientNet-b2) with frozen pre-trained encoders for each task

Metrics

- Semantic Segmentation metrics:
  - task-specific Loss
  - mean Intersection over Union
  - Pixel Accuracy
- Depth Estimation metrics:
  - task-specific Loss
  - Absolute Error
  - Relative Error
- Surface Normals Prediction metrics:
  - task-specific Loss
  - mean Angle Distance
  - median Angle Distance
  - mean of Angle Errors within 11.25°
  - mean of Angle Errors within 22.5°
  - mean of Angle Errors within 30°

Models' Parameter Count

Parameters	**Architecture**					
	EfficientPS-SS	EfficientPS-DE	EfficientPS-SNP	MT-Net DWA	**MBMT-Net DWA**	MTAN DWA
Trainable	10094853	11268345	11268859	32654444	**8722196**	44229076
Total	10094853	11268345	11268859	32654444	**32063398**	44229076

Table: Analysis of the number of parameters for the architectures involved in the experiments

MBMT-Net Performance Compared to Other Models

Task	Measure	Architecture			
		EfficientPS	MT-Net DWA	**MBMT-Net DWA**	MTAN DWA (SOTA)
Semantic Segmentation	**Loss (↓)**	**2.1877**	**2.1990**	**2.1605**	-
	mIoU (↑)	21.23	20.87	**24.84**	17.15
	PAcc (↑)	50.06	48.88	52.76	**54.97**
Depth Estimation	**Loss (↓)**	**0.6861**	0.6185	**0.6076**	-
	Abs (↓)	0.6861	0.6185	0.6076	**0.5906**
	Rel (↓)	0.2913	0.2646	**0.2569**	**0.2569**
Surface Normal Prediction	**Loss (↓)**	**0.2249**	**0.2358**	**0.2174**	-
	Mean (↓)	32.8582	33.9981	32.0937	**31.60**
	Med (↓)	27.2290	28.7907	26.3163	**25.46**
	<11.25 (↑)	20.49	18.81	21.69	**22.48**
	<22.5 (↑)	42.22	39.75	43.66	**44.86**
	<30 (↑)	54.22	51.88	55.62	**57.24**

Table: Results of the previously mentioned architectures compared to state-of-the-art MTAN [1] on the three tasks of choice

Relative Performance Improvements of MBMT-Net

Improvement OURS									Angle Distance		Within t°		
Trainable #P (↓)	Total #P (↓)	SEMANTIC_LOSS (↓)	MEAN_IOU (↑)	PIX_ACC (↑)	DEPTH_LOSS (↓)	ABS_ERR (↓)	REL_ERR (↓)	NORMAL_LOSS (↓)	MEAN (↓)	MED (↓)	<11.25 (↑)	<22.5 (↑)	<30 (↑)
vs MTAN (SOTA)													
80.28%	26.32%		44.84%	-4.02%		-2,.8%	0.00%		-1.56%	-3.36%	-3.51%	-2.67%	-2.83%
vs EfficientPS													
		1.24%	17.00%	5.39%	11.44%	11.44%	11.81%	3.33%	2.33%	3.35%	5.86%	3.41%	2.58%
vs MT-Net													
73.29%	0.20%	1.75%	19.02%	7.94%	1.76%	1.76%	2.91%	7.80%	5.60%	8.59%	15.31%	9.84%	7.21%

Table: Relative improvements of MBMT-Net considering all the proposed metrics

Conclusions and Future Enhancements

Conclusions:

- Having the same number of parameters, the MB-encoder outperforms other methods
- MBMT-Net reaches state-of-the-art performance, while having 12.16M less than MTAN
- More efficient training schedule when using pre-trained backbones

Future Enhancements:

- Testing multiple ST architectures to prove MBMT-Net's effectiveness
- Potential interest increase in the Distributed AI paradigm

Thank you!

Questions?

Bibliography I

- S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 1871–1880, Computer Vision Foundation / IEEE, 2019.
- P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.
- R. Mohan and A. Valada, "Efficientps: Efficient panoptic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1551–1579, 2021.

Bibliography II


M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114, PMLR, 2019.


J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255, IEEE Computer Society, 2009.

Bibliography III

T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 936–944, IEEE Computer Society, 2017.

L. Chen, M. D. Collins, Y. Zhu, G. Papandreou, B. Zoph, F. Schroff, H. Adam, and J. Shlens, "Searching for efficient multi-scale architectures for dense image prediction," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada* (S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), pp. 8713–8724, 2018.

Bibliography IV


Y. Liu, Y. Wang, S. Wang, T. Liang, Q. Zhao, Z. Tang, and H. Ling, "Cbnet: A novel composite backbone network architecture for object detection," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 11653–11660, AAAI Press, 2020.


T. Liang, X. Chu, Y. Liu, Y. Wang, Z. Tang, W. Chu, J. Chen, and H. Ling, "Cbnetv2: A composite backbone network architecture for object detection," *CoRR*, vol. abs/2107.00420, 2021.

Bibliography V


A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, eds.), pp. 8024–8035, 2019.