Mitigating Task Interference in Multi-Task Learning via Explicit Task Routing with Non-Learnable Primitives (Supplementary Materials)

Chuntao Ding[1*] Zhichao Lu[2†] Shangguang Wang[3] Ran Cheng[4] Vishnu N. Boddeti[5]
[1] Beijing Jiaotong University [2] Sun Yat-sen University [3] Beijing University of Posts and Telecommunications
[4] Southern University of Science and Technology [5] Michigan State University
chuntaoding@163.com {luzhichaocn, ranchengcn}@gmail.com sgwang@bupt.edu.cn vishnu@msu.edu

This appendix includes the following:

1. Extended description of Related Work in Section A.
2. Additional results of configurations on NLPs-based feature extraction in Section B.
3. Extended description of datasets and baseline methods in Section C.

A. An Extended Description of Related Work

Existing methods related to MTL architectures can be classified into encoder or decoder-focused ones. Encoder-focused approaches primarily lay emphasis on architectures that can encode multi-purpose feature representations through supervision from multiple tasks. Such encoding is typically achieved, for example, via feature fusion [5,12,14], branching [7,10,11,19], self-supervision [3], shared and task-specific modules [8,13], filter grouping [1], filter modulation [6, 27], task routing [12, 16, 18], or neural architecture search [4]. Decoder-focused approaches start from the feature representations learned at the encoding stage, and further refine them at the decoding stage by distilling information across tasks in a one-off [22], sequential [24], recursive [25], or even multi-scale [20] manner. Due to inherent layer sharing, the approaches above typically suffer from task interference and negative transfer [21].

In the context of MTL, our explicit task routing layer is conceptually related to [8, 13], but notably different in motivation and design. First, both of these two approaches operate on the features obtained from a shared backbone to extract task-specific features. In contrast, our shared and task-specific branches operate in parallel and extract features from the common features extracted by the non-learnable layer. Second, these existing approaches utilize attention mechanisms to distill task-specific features from the shared features, while we use lightweight 1×1 convolution for the same purpose. Third, our explicit task routing layer is tailored to exploit the non-learnable layer for MTL optimally. Finally, unlike baselines, our multi-branch design affords simple and explicit control over the ratio of shared and task-specific parameters.

Additionally, our work is also closely related to reparameterized convolutions for multi-task learning (RCM) [6], which first introduced the concept of using non-learnable convolutional filters for MTL. However, there are three notable differences. First, the non-learnable layer of RCM only includes standard convolution, while we consider other non-learnable primitives such as pooling, identity, and additive noise [23] operations. Second, RCM uses pre-trained network weights to initialize non-learnable convolutional filters, while in our case they are sampled from a random distribution. Relying on pre-trained weights limits RCM's ability to reduce the model size and its generalizability to architectures without readily available pre-trained weights. Finally, there is no collaboration between tasks in RCM as it only comprises task-specific modulators, while we utilize a shared branch to help tasks use each other's training signals. Having both shared and task-specific branches allows tasks to amortize parameters that are commonly useful across multiple tasks, thereby minimizing redundancy in the task-specific branches, unlike RCM. Moreover, our method also offers fine-grained control over the ratio of parameters that are shared or task-specific.

B. Results of NLPs-based feature extraction

In this section, we first provide the full version of Table 1 from the main paper in Table S1, showing the effect of different configurations of NLPs on CelebA multi-attribute classification. Then, we present the effect of different hyperparameter settings of NLPs on NYU-v2 dense prediction MTL problem in Figure S1.

* *Work done as a visiting scholar at Michigan State University.*
† *Corresponding author*

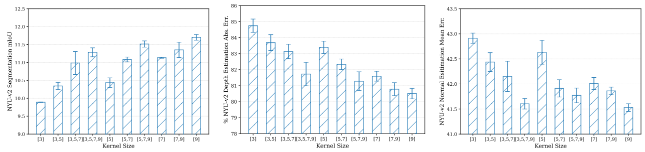
Table S1. Effect of different configurations of NLPs on CelebA multi-attribute classification.

Non-learnable Operators					CelebA		
Avg. pool	Max pool	Conv	Shift	Noise	Precision	Recall	F-Score
		✓			$73.37_{\pm 0.19}$	$59.00_{\pm 0.11}$	$61.08_{\pm 0.17}$
	✓	✓			$72.50_{\pm 0.13}$	$57.87_{\pm 0.09}$	$61.10_{\pm 0.12}$
	✓				$72.29_{\pm 0.14}$	$58.65_{\pm 0.27}$	$61.30_{\pm 0.04}$
✓	✓	✓			$73.24_{\pm 0.17}$	$58.87_{\pm 0.30}$	$61.46_{\pm 0.13}$
✓	✓	✓	✓		$73.37_{\pm 0.31}$	$58.67_{\pm 0.07}$	$61.56_{\pm 0.24}$
✓	✓				$72.92_{\pm 0.13}$	$58.76_{\pm 0.08}$	$61.57_{\pm 0.06}$
✓	✓		✓		$73.51_{\pm 0.39}$	$59.47_{\pm 0.25}$	$61.81_{\pm 0.15}$
	✓		✓		$72.80_{\pm 0.33}$	$59.39_{\pm 0.13}$	$61.83_{\pm 0.20}$
	✓	✓	✓		$72.93_{\pm 0.15}$	$59.47_{\pm 0.15}$	$62.04_{\pm 0.20}$
			✓		$73.56_{\pm 0.18}$	$59.90_{\pm 0.14}$	$62.16_{\pm 0.07}$
		✓	✓		$73.52_{\pm 0.62}$	$59.16_{\pm 0.08}$	$62.19_{\pm 0.24}$
✓			✓		$74.23_{\pm 0.48}$	$59.68_{\pm 0.17}$	$62.36_{\pm 0.13}$
✓		✓	✓		$73.97_{\pm 0.47}$	$59.85_{\pm 0.17}$	$62.41_{\pm 0.08}$
✓		✓			$74.81_{\pm 0.39}$	$59.84_{\pm 0.29}$	$62.84_{\pm 0.11}$
✓					$74.75_{\pm 0.36}$	$60.49_{\pm 0.05}$	$63.03_{\pm 0.32}$
✓	✓	✓	✓	✓	$74.56_{\pm 0.24}$	$61.24_{\pm 0.61}$	$64.08_{\pm 0.21}$
✓	✓	✓		✓	$74.38_{\pm 0.10}$	$61.13_{\pm 0.10}$	$64.14_{\pm 0.02}$
✓	✓		✓	✓	$74.24_{\pm 0.08}$	$61.14_{\pm 0.29}$	$64.21_{\pm 0.09}$
	✓	✓	✓	✓	$74.50_{\pm 0.30}$	$61.17_{\pm 0.46}$	$64.32_{\pm 0.04}$
	✓		✓	✓	$74.40_{\pm 0.32}$	$62.07_{\pm 0.19}$	$64.51_{\pm 0.22}$
				✓	$75.67_{\pm 0.25}$	$59.74_{\pm 0.33}$	$64.54_{\pm 0.21}$
✓	✓			✓	$74.35_{\pm 0.03}$	$62.18_{\pm 0.16}$	$64.61_{\pm 0.15}$
	✓	✓		✓	$74.91_{\pm 0.14}$	$62.16_{\pm 0.34}$	$64.65_{\pm 0.05}$
✓		✓	✓	✓	$75.35_{\pm 0.10}$	$61.99_{\pm 0.20}$	$65.03_{\pm 0.02}$
	✓			✓	$74.61_{\pm 0.14}$	$61.93_{\pm 0.23}$	$65.07_{\pm 0.05}$
✓			✓	✓	$75.31_{\pm 0.22}$	$62.89_{\pm 0.21}$	$65.42_{\pm 0.15}$
		✓	✓	✓	$75.50_{\pm 0.15}$	$62.70_{\pm 0.41}$	$65.65_{\pm 0.16}$
			✓	✓	$75.74_{\pm 0.29}$	$62.80_{\pm 0.19}$	$65.81_{\pm 0.13}$
✓				✓	$75.72_{\pm 0.08}$	$63.19_{\pm 0.39}$	$66.08_{\pm 0.10}$
✓		✓		✓	$75.82_{\pm 0.29}$	$63.19_{\pm 0.36}$	$66.26_{\pm 0.25}$
		✓		✓	$76.29_{\pm 0.25}$	$62.47_{\pm 0.60}$	$66.40_{\pm 0.32}$
Standard learnable convolution					$67.67_{\pm 0.75}$	$59.84_{\pm 0.33}$	$62.86_{\pm 0.07}$

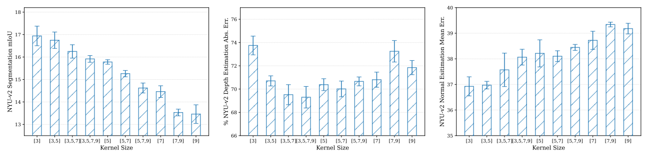
C. Description of Datasets and Baselines

In this section, we first provide the additional details of CelebA and Cityscapes datasets in Table 1 and 2, respectively. Then, we provide a brief overview of the baseline methods that we compared against in this work, as follows:

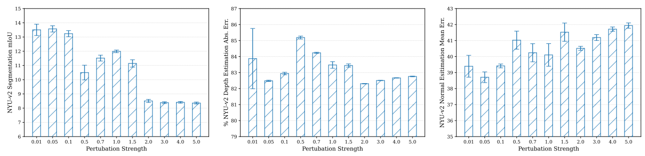
- STL: single task learning with one network for each task.
- Hard sharing: standard multi-task learning, i.e., a fully shared network with uniform task weighting.
- GradNorm [26]: a MTL method with a fully shared network and learnable tasks weighting.
- MGDA-UB [17]: a multi-objective alternative to MTL with a fully shared network.
- Task Routing [18]: a parameter partitioning method with randomly initialized binary masks.
- Max. Roaming [15]: another parameter partitioning method with dynamic masks.
- Cross-stitch [14]: a soft-sharing method with feature fusion.
- MTAN [8]: another soft-sharing method with attention.



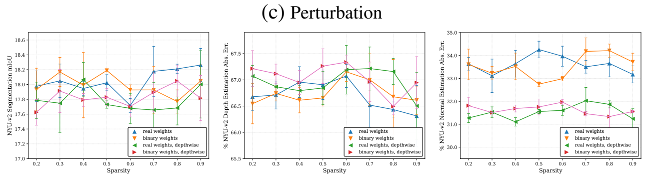
(a) Avg. pooling



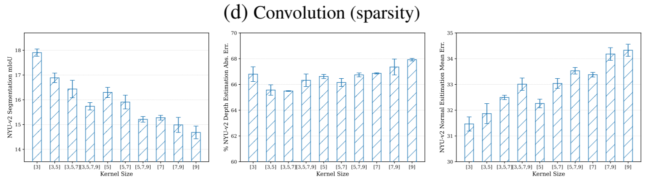
(b) Max pooling



(c) Perturbation



(d) Convolution (sparsity)



(e) Convolution (kernel size)

Figure S1. Effect of different hyperparameters of individual NLP on NYU-v2 dense prediction MTL problem. For each sub-figure (a) - (e), we show the semantic segmentation mIoU (↑), depth estimation absolute error (↓), and surface normal estimation mean error (↓).

References

[1] Felix JS Bragman, Ryutaro Tanno, Sebastien Ourselin, Daniel C Alexander, and Jorge Cardoso. Stochastic filter groups for multi-task cnns: Learning specialist and generalist convolution kernels. In *International Conference on Computer Vision (ICCV)*, pages 1385–1394, 2019. 1

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 3

[3] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2070–2079, 2017. 1

[4] Yuan Gao, Haoping Bai, Zequn Jie, Jiayi Ma, Kui Jia, and Wei Liu. Mtl-nas: Task-agnostic neural architecture

Table 1. Details of the 40 attributes provided by the CelebA dataset [9]. For visualization purposes, we group them into eight categories.

Group	40-attribute
global	attractive, blurry, chubby, double chin, heavy makeup, male, oval face, pale skin, young
eyes	bags under eyes, eyeglasses, narrow eyes, arched eyebrows, bushy eyebrows
hair	bald, bangs, black hair, blond hair, brown hair, gray hair, receding hairline, straight hair, wavy hair
mouth	big lips, mouth slightly open, smiling, wearing lipstick
nose	big nose, pointy nose
beard	5 o'clock shadow, goatee, mustache, no beard, sideburns
cheek	high cheekbones, rosy cheeks
wearings	wearing earrings, wearing hat, wearing necklace, wearing necktie

Table 2. Three levels of semantic categories for the Cityscapes dataset [2, 8]. The experimental results in this work are based on the 7-class setting.

2-class	7-class	19-class
background	void	void
	flat	road, sidewalk
	construction	building, wall, fence
	object	pole, traffic light, traffic sign
	nature	vegetation, terrain
	sky	sky
foreground	human	person, rider
	vehicle	carm truck, bus, caravan, trailer, train, motorcycle

search towards general-purpose multi-task learning. In *IEEE Conference on Computer Vision and pattern Recognition (CVPR)*, pages 11540–11549, 2020. 1

[5] 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 *IEEE Conference on Computer Vision and pattern Recognition (CVPR)*, pages 3205–3214, 2019. 1

[6] Menelaos Kanakis, David Bruggemann, Suman Saha, Stamatios Georgoulis, Anton Obukhov, and Luc Van Gool. Reparameterizing convolutions for incremental multi-task learning without task interference. In *European Conference on Computer Vision (ECCV)*, pages 689–707, 2020. 1

[7] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5454–5463, 2017. 1

[8] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1871–1880, 2019. 1, 2, 3

[9] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 3

[10] Mingsheng Long and Jianmin Wang. Learning multiple tasks with deep relationship networks, 2015. In CoRR abs/1506.02117. 1

[11] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Schmidt Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *IEEE Conference on Computer Vision and pattern Recognition (CVPR)*, pages 1131–1140, 2017. 1

[12] Jiaqi Ma, Zhe Zhao, Jilin Chen, Ang Li, Lichan Hong, and Ed H. Chi. Snr: Sub-network routing for flexible parameter sharing in multi-task learning. In *Association for the Advance of Artificial Intelligence (AAAI)*, pages 216–223, 2019. 1

[13] Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1851–1860, 2019. 1

[14] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *IEEE Conference on Computer Vision and pattern Recognition (CVPR)*, pages 3994–4003, 2016. 1, 2

[15] Lucas Pascal, Pietro Michiardi, Xavier Bost, Benoit Huet, and Maria A Zuluaga. Maximum roaming multi-task learning. In *AAAI Conference on Artificial Intelligence*, pages 9331–9341, 2021. 2

[16] Clemens Rosenbaum, Tim Klinger, and Matthew Riemer. Routing networks: Adaptive selection of non-linear functions for multi-task learning. In *International Conference on Learning Representations (ICLR)*, 2018. 1

[17] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 525–536, 2018. 2

[18] Gjorgji Strezoski, Nanne van Noord, and Marcel Worring. Many task learning with task routing. In *International Conference on Computer Vision (ICCV)*, pages 1375–1384, 2019. 1, 2

[19] Simon Vandenhende, Bert De Brabandere, and Luc Van Gool. Branched multi-task networks: Deciding what layers to share, 2019. In *British Machine Vision Conference (BMVC)*. 1

[20] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *European Conference on Computer Vision (ECCV)*, pages 527–543, 2020. 1

[21] Sen Wu, Hongyang R. Zhang, and Christopher Re. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations (ICLR)*, 2020. 1

[22] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing.

In *IEEE Conference on Computer Vision and pattern Recognition (CVPR)*, pages 675–684, 2018. 1

[23] Felix Juefei Xu, Vishnu Naresh Boddeti, and Marios Savvides. Perturbative neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3310–3318, 2018. 1

[24] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *European Conference on Computer Vision (ECCV)*, pages 235–251, 2018. 1

[25] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *IEEE Conference on Computer Vision and pattern Recognition (CVPR)*, pages 4106–4115, 2019. 1

[26] Chen Zhao, Badrinarayanan Vijay, Lee Chen-Yu, and Rabinovich Andrew. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning (ICML)*, pages 793–802, 2018. 2

[27] Xiangyun Zhao, Haoxiang Li, Xiaohui Shen, Xiaodan Liang, and Ying Wu. A modulation module for multi-task learning with applications in image retrieval. In *European Conference on Computer Vision (ECCV)*, pages 415–432, 2018. 1