In this document, we provide the results of different combinations of different optimization strategies and architectures on the *NYUv2* dataset. There are something required special attention as follows.

- We comment out the code about reproducibility (`https://github.com/median-research-group/LibMTL/blob/main/LibMTL/utils.py#L18-L20`) for faster running speed;
- Each experiment is run only once here;
- We only use the default hyperparameters for each method, refer to Table 1 for details;
- All experiments are conduced on the *NYUv2* dataset only.
- A few experiments have appeared the Out Of Memory (OOM) error because of the insufficient GPU memory.

Table 1: Hyperparameters Configuration.

| | **Configuration** |
|---|---|
| common | `GPU`: NVIDIA GeForce RTX 3090<br>`multi_input`: False<br>`aug`: False; `seed`: 0<br>`train_bs`: 8; `test_bs`: 8; epochs: 200<br>`optim`: Adam; `lr`: 0.0001; `weight_decay`: 1e-05<br>`scheduler`: step; `step_size`: 100; `gamma`: 0.5 |
| GradNorm | `rep_grad`: False; `alpha`: 1.5 |
| MGDA | `rep_grad`: True; `mgda_gn`: none |
| DWA | `T`: 2 |
| GradDrop | `leak`: 0.0 |
| IMTL | `rep_grad`: True |
| GradVac | `beta`: 0.5 |
| CAGrad | `calpha`: 0.5; `rescale`: 1 |
| Nash-MTL | `update_weights_every`: 1<br>`optim_niter`: 20; `max_norm`: 1.0 |
| MMoE | `img_size`: [3, 288, 384]<br>`num_experts`: [2] |
| CGC | `img_size`: [3, 288, 384]<br>`num_experts`: [1, 1, 1, 1] |
| PLE | `img_size`: [3, 288, 384]<br>`num_experts`: [1, 1, 1, 1]<br>`train_bs`: 4; `test_bs`: 4 |
| DSelect-k | `img_size`: [3, 288, 384]<br>`num_experts`: [2]<br>`kgamma`: 1.0; `num_nonzeros`: 2 |

Table 2: Performance on the *NYUv2* dataset with 3 tasks on **HPS** [1] architecture.

| | **Segmentation** | | **Depth** | | **Normal** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | **Angle Distance** | | **Within $t^\circ$** | | |
| | **mIoU↑** | **PAcc↑** | **AErr↓** | **RErr↓** | **Mean↓** | **MED↓** | **11.25↑** | **22.5↑** | **30↑** |
| EW | 54.10 | 75.78 | 0.3771 | 0.1596 | 23.53 | 17.20 | 34.55 | 60.75 | 72.06 |
| GradNorm [2] | 53.57 | 75.21 | 0.3853 | 0.1623 | 23.32 | 16.62 | 35.81 | 61.85 | 72.73 |
| UW [7] | 54.02 | 75.66 | 0.3838 | 0.1573 | 23.45 | 16.86 | 35.43 | 61.32 | 72.33 |
| MGDA [15] | 50.15 | 73.44 | 0.3839 | 0.1563 | 22.78 | 16.05 | 37.04 | 63.07 | 73.65 |
| DWA [11] | 53.55 | 75.42 | 0.3811 | 0.1589 | 23.63 | 17.15 | 34.78 | 60.71 | 71.87 |
| GLS [4] | 54.46 | 76.01 | 0.3793 | 0.1566 | 22.73 | 16.16 | 36.76 | 62.94 | 73.68 |
| PCGrad [18] | 54.03 | 75.66 | 0.3814 | 0.1570 | 23.62 | 17.01 | 35.10 | 60.96 | 71.97 |
| GradDrop [3] | 53.90 | 75.62 | 0.3851 | 0.1580 | 23.67 | 17.04 | 34.99 | 60.88 | 71.91 |
| IMTL [10] | 53.53 | 75.46 | 0.3842 | 0.1578 | 22.99 | 16.44 | 36.16 | 62.28 | 73.20 |
| GradVac [17] | 53.99 | 75.74 | 0.3842 | 0.1602 | 23.51 | 16.93 | 35.26 | 61.14 | 72.17 |
| CAGrad [9] | 53.77 | 75.80 | 0.3836 | 0.1584 | 22.53 | 15.86 | 37.40 | 63.58 | 74.20 |
| Nash-MTL [14] | 53.38 | 75.01 | 0.3814 | 0.1570 | 22.61 | 15.87 | 37.30 | 63.68 | 74.28 |
| RLW [8] | 54.52 | 75.78 | 0.3890 | 0.1566 | 23.34 | 16.83 | 35.45 | 61.43 | 72.45 |

Table 3: Performance on the *NYUv2* dataset with 3 tasks on **Cross-stitch** [13] architecture.

| | **Segmentation** | | **Depth** | | **Normal** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | **Angle Distance** | | **Within $t^\circ$** | | |
| | **mIoU↑** | **PAcc↑** | **AErr↓** | **RErr↓** | **Mean↓** | **MED↓** | **11.25↑** | **22.5↑** | **30↑** |
| EW | 53.14 | 75.24 | 0.3824 | 0.1621 | 23.03 | 16.29 | 36.77 | 62.30 | 72.98 |
| GradNorm [2] | 53.70 | 75.30 | 0.3790 | 0.1591 | 23.12 | 16.50 | 36.30 | 61.95 | 72.77 |
| UW [7] | 53.39 | 75.37 | 0.3780 | 0.1583 | 23.12 | 16.32 | 36.77 | 62.29 | 72.94 |
| MGDA [15] | 53.78 | 75.50 | 0.3758 | 0.1560 | 22.70 | 15.91 | 37.53 | 63.21 | 73.72 |
| DWA [11] | 52.91 | 75.06 | 0.3795 | 0.1577 | 23.19 | 16.42 | 36.61 | 62.02 | 72.72 |
| GLS [4] | 53.50 | 75.46 | 0.3798 | 0.1576 | 22.52 | 15.73 | 37.90 | 63.60 | 74.02 |
| PCGrad [18] | 52.71 | 75.10 | 0.3824 | 0.1607 | 22.97 | 16.14 | 37.08 | 62.66 | 73.19 |
| GradDrop [3] | | | | | OOM | | | | |
| IMTL [10] | 53.49 | 75.65 | 0.3770 | 0.1565 | 22.81 | 15.96 | 37.41 | 63.13 | 73.65 |
| GradVac [17] | 53.45 | 75.22 | 0.3787 | 0.1570 | 23.25 | 16.54 | 36.34 | 61.76 | 72.51 |
| CAGrad [9] | 53.69 | 75.17 | 0.3757 | 0.1503 | 22.45 | 15.80 | 37.71 | 63.52 | 74.08 |
| Nash-MTL [14] | 53.17 | 75.12 | 0.3701 | 0.1542 | 22.63 | 15.81 | 37.62 | 63.48 | 73.98 |
| RLW [8] | 53.27 | 75.17 | 0.3772 | 0.1546 | 22.65 | 15.84 | 37.58 | 63.37 | 73.81 |

Table 4: Performance on the *NYUv2* dataset with 3 tasks on **MMoE** [12] architecture.

| | **Segmentation** | | **Depth** | | **Normal** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | **Angle Distance** | | **Within $t^{\circ}$** | | |
| | **mIoU↑** | **PAcc↑** | **AErr↓** | **RErr↓** | **Mean↓** | **MED↓** | **11.25↑** | **22.5↑** | **30↑** |
| EW | 53.02 | 75.09 | 0.3884 | 0.1631 | 22.96 | 16.37 | 36.31 | 62.46 | 73.30 |
| GradNorm [2] | 53.09 | 75.07 | 0.3946 | 0.1698 | 22.93 | 16.23 | 36.81 | 62.59 | 73.32 |
| UW [7] | 53.49 | 75.38 | 0.3813 | 0.1605 | 22.76 | 15.99 | 37.23 | 63.16 | 73.74 |
| MGDA [15] | 50.69 | 73.82 | 0.3772 | 0.1574 | 22.78 | 16.22 | 36.68 | 62.70 | 73.43 |
| DWA [11] | 53.18 | 74.70 | 0.3859 | 0.1613 | 23.25 | 16.61 | 35.87 | 61.90 | 72.76 |
| GLS [4] | 53.35 | 75.28 | 0.3778 | 0.1576 | 22.54 | 15.77 | 37.70 | 63.65 | 74.12 |
| PCGrad [18] | 54.24 | 75.68 | 0.3832 | 0.1601 | 22.58 | 15.88 | 37.42 | 63.45 | 73.98 |
| GradDrop [3] | | | | | OOM | | | | |
| IMTL [10] | 53.83 | 75.59 | 0.3788 | 0.1601 | 22.79 | 16.16 | 36.88 | 62.78 | 73.46 |
| GradVac [17] | 52.93 | 75.04 | 0.3903 | 0.1565 | 22.89 | 16.04 | 37.00 | 63.03 | 73.57 |
| CAGrad [9] | 53.15 | 74.96 | 0.3933 | 0.1650 | 22.38 | 15.66 | 37.93 | 63.88 | 74.36 |
| Nash-MTL [14] | 53.67 | 75.62 | 0.3885 | 0.1604 | 22.28 | 15.76 | 37.61 | 63.86 | 74.49 |
| RLW [8] | 53.10 | 74.77 | 0.3955 | 0.1696 | 22.32 | 15.51 | 38.29 | 64.15 | 74.42 |

Table 5: Performance on the *NYUv2* dataset with 3 tasks on **MTAN** [11] architecture.

| | **Segmentation** | | **Depth** | | **Normal** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | **Angle Distance** | | **Within $t^{\circ}$** | | |
| | **mIoU↑** | **PAcc↑** | **AErr↓** | **RErr↓** | **Mean↓** | **MED↓** | **11.25↑** | **22.5↑** | **30↑** |
| EW | 54.38 | 75.84 | 0.3764 | 0.1561 | 22.98 | 16.36 | 36.35 | 62.44 | 73.27 |
| GradNorm [2] | 54.01 | 75.67 | 0.3838 | 0.1554 | 23.25 | 16.59 | 35.97 | 61.89 | 72.82 |
| UW [7] | 54.01 | 75.75 | 0.3729 | 0.1515 | 23.04 | 16.25 | 36.79 | 62.53 | 73.21 |
| MGDA [15] | 51.43 | 74.22 | 0.3712 | 0.1508 | 22.20 | 15.30 | 38.79 | 64.73 | 74.89 |
| DWA [11] | 54.82 | 75.93 | 0.3780 | 0.1582 | 22.94 | 16.34 | 36.53 | 62.44 | 73.21 |
| GLS [4] | 55.16 | 76.34 | 0.3726 | 0.1551 | 22.58 | 15.82 | 37.63 | 63.59 | 74.12 |
| PCGrad [18] | 54.63 | 75.95 | 0.3810 | 0.1585 | 23.48 | 17.02 | 35.04 | 60.99 | 72.16 |
| GradDrop [3] | | | | | OOM | | | | |
| IMTL [10] | 54.35 | 75.84 | 0.3722 | 0.1553 | 22.45 | 15.72 | 37.76 | 63.85 | 74.34 |
| GradVac [17] | 53.65 | 75.59 | 0.3821 | 0.1612 | 23.43 | 17.01 | 35.12 | 61.04 | 72.21 |
| CAGrad [9] | 53.87 | 75.40 | 0.3815 | 0.1557 | 22.55 | 15.91 | 37.36 | 63.43 | 74.08 |
| Nash-MTL [14] | 54.78 | 75.85 | 0.3788 | 0.1546 | 22.46 | 15.60 | 38.19 | 64.01 | 74.41 |
| RLW [8] | 55.19 | 76.13 | 0.3758 | 0.1568 | 22.79 | 16.05 | 37.11 | 63.08 | 73.70 |

Table 6: Performance on the *NYUv2* dataset with 3 tasks on **CGC** [16] architecture.

| | Segmentation | | Depth | | Normal | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Angle Distance | | Within $t^\circ$ | | |
| | mIoU↑ | PAcc↑ | AErr↓ | RErr↓ | Mean↓ | MED↓ | 11.25↑ | 22.5↑ | 30↑ |
| EW | 53.40 | 75.07 | 0.3897 | 0.1650 | 22.17 | 15.41 | 38.50 | 64.47 | 74.76 |
| GradNorm [2] | 53.01 | 75.07 | 0.3957 | 0.1653 | 22.17 | 15.42 | 38.45 | 64.51 | 74.81 |
| UW [7] | 52.78 | 75.08 | 0.3832 | 0.1559 | 22.35 | 15.26 | 38.79 | 64.58 | 74.61 |
| MGDA [15] | 54.34 | 76.20 | 0.3799 | 0.1573 | 22.09 | 15.41 | 38.39 | 64.63 | 75.03 |
| DWA [11] | 52.35 | 74.90 | 0.3992 | 0.1694 | 22.09 | 15.35 | 38.74 | 64.55 | 74.81 |
| GLS [4] | 53.36 | 75.38 | 0.3829 | 0.1595 | 22.12 | 15.44 | 38.42 | 64.50 | 74.85 |
| PCGrad [18] | 52.90 | 74.98 | 0.3968 | 0.1646 | 22.17 | 15.51 | 38.25 | 64.27 | 74.63 |
| GradDrop [3] | | | | | OOM | | | | |
| IMTL [10] | 52.77 | 75.23 | 0.3920 | 0.1593 | 22.71 | 15.90 | 37.43 | 63.41 | 73.96 |
| GradVac [17] | 53.06 | 75.01 | 0.4000 | 0.1666 | 22.22 | 15.40 | 38.34 | 64.58 | 74.77 |
| CAGrad [9] | 53.61 | 74.94 | 0.3952 | 0.1644 | 22.18 | 15.47 | 38.26 | 64.41 | 74.77 |
| Nash-MTL [14] | | | | | OOM | | | | |
| RLW [8] | 53.42 | 75.21 | 0.3916 | 0.1646 | 21.78 | 15.09 | 39.21 | 65.22 | 75.43 |

Table 7: Performance on the *NYUv2* dataset with 3 tasks on **PLE** [16] architecture.

| | Segmentation | | Depth | | Normal | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Angle Distance | | Within $t^\circ$ | | |
| | mIoU↑ | PAcc↑ | AErr↓ | RErr↓ | Mean↓ | MED↓ | 11.25↑ | 22.5↑ | 30↑ |
| EW | 53.22 | 74.89 | 0.4002 | 0.1658 | 22.04 | 15.17 | 38.75 | 65.24 | 75.42 |
| GradNorm [2] | 53.17 | 74.98 | 0.3809 | 0.1601 | 21.96 | 15.12 | 39.25 | 65.05 | 75.18 |
| UW [7] | 53.13 | 74.89 | 0.3767 | 0.1536 | 22.27 | 15.39 | 38.52 | 64.47 | 74.72 |
| MGDA [15] | 53.55 | 75.42 | 0.3985 | 0.1663 | 22.07 | 15.35 | 38.58 | 64.73 | 75.02 |
| DWA [11] | 53.06 | 75.00 | 0.3850 | 0.1561 | 21.97 | 15.21 | 38.92 | 64.98 | 75.25 |
| GLS [4] | 53.28 | 75.33 | 0.3736 | 0.1549 | 22.14 | 15.49 | 38.43 | 64.22 | 74.58 |
| PCGrad [18] | 52.99 | 74.93 | 0.3848 | 0.1592 | 22.55 | 15.56 | 38.29 | 63.90 | 74.17 |
| GradDrop [3] | | | | | OOM | | | | |
| IMTL [10] | 53.48 | 75.58 | 0.3876 | 0.1564 | 21.79 | 15.04 | 39.28 | 65.43 | 75.59 |
| GradVac [17] | 53.03 | 74.96 | 0.3926 | 0.1594 | 22.02 | 15.36 | 38.72 | 64.52 | 74.86 |
| CAGrad [9] | 52.76 | 75.11 | 0.3923 | 0.1586 | 22.12 | 15.58 | 38.14 | 64.16 | 74.66 |
| Nash-MTL [14] | 50.80 | 73.84 | 0.3813 | 0.1562 | 22.28 | 15.57 | 37.90 | 64.30 | 74.78 |
| RLW [8] | 52.24 | 74.08 | 0.3863 | 0.1621 | 22.26 | 15.39 | 38.78 | 64.24 | 74.46 |

Table 8: Performance on the *NYUv2* dataset with 3 tasks on **LTB** [5] architecture.

| | **Segmentation** | | **Depth** | | **Normal** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | **Angle Distance** | | **Within $t^\circ$** | | |
| | **mIoU↑** | **PAcc↑** | **AErr↓** | **RErr↓** | **Mean↓** | **MED↓** | **11.25↑** | **22.5↑** | **30↑** |
| EW | 52.36 | 74.75 | 0.3850 | 0.1611 | 23.26 | 16.60 | 36.21 | 61.69 | 72.47 |
| GradNorm [2] | 51.93 | 74.60 | 0.3854 | 0.1618 | 23.25 | 16.54 | 36.27 | 61.82 | 72.55 |
| UW [7] | 52.19 | 74.68 | 0.3886 | 0.1576 | 23.32 | 16.47 | 36.46 | 61.81 | 72.42 |
| MGDA [15] | 52.85 | 74.68 | 0.3849 | 0.1585 | 22.97 | 16.13 | 37.11 | 62.67 | 73.18 |
| DWA [11] | 52.25 | 74.50 | 0.3865 | 0.1579 | 23.39 | 16.70 | 35.95 | 61.45 | 72.21 |
| GLS [4] | 52.97 | 74.88 | 0.3797 | 0.1536 | 22.90 | 15.98 | 37.44 | 62.93 | 73.34 |
| PCGrad [18] | 52.40 | 74.83 | 0.3847 | 0.1600 | 23.26 | 16.58 | 36.17 | 61.76 | 72.55 |
| GradDrop [3] | | | | | OOM | | | | |
| IMTL [10] | 52.74 | 75.08 | 0.3783 | 0.1579 | 22.97 | 16.21 | 36.91 | 62.57 | 73.22 |
| GradVac [17] | 52.35 | 74.81 | 0.3817 | 0.1593 | 23.36 | 16.62 | 36.10 | 61.59 | 72.33 |
| CAGrad [9] | 53.25 | 75.36 | 0.3711 | 0.1567 | 22.65 | 15.94 | 37.42 | 63.22 | 73.79 |
| Nash-MTL [14] | | | | | OOM | | | | |
| RLW [8] | 52.62 | 74.60 | 0.3829 | 0.1622 | 22.95 | 15.94 | 37.67 | 62.85 | 73.16 |

Table 9: Performance on the *NYUv2* dataset with 3 tasks on **DSelect-k** [6] architecture.

| | **Segmentation** | | **Depth** | | **Normal** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | **Angle Distance** | | **Within $t^\circ$** | | |
| | **mIoU↑** | **PAcc↑** | **AErr↓** | **RErr↓** | **Mean↓** | **MED↓** | **11.25↑** | **22.5↑** | **30↑** |
| EW | 54.03 | 75.37 | 0.3854 | 0.1597 | 23.45 | 16.88 | 35.37 | 61.23 | 72.22 |
| GradNorm [2] | 53.54 | 75.43 | 0.3845 | 0.1625 | 23.05 | 16.46 | 36.19 | 62.22 | 73.07 |
| UW [7] | 54.34 | 75.55 | 0.3796 | 0.1587 | 23.17 | 16.55 | 36.07 | 61.96 | 72.79 |
| MGDA [15] | 53.88 | 75.72 | 0.3874 | 0.1613 | 22.25 | 15.63 | 37.99 | 63.99 | 74.48 |
| DWA [11] | 53.59 | 75.02 | 0.3849 | 0.1621 | 22.51 | 15.32 | 38.82 | 64.15 | 74.09 |
| GLS [4] | 54.76 | 75.96 | 0.3801 | 0.1569 | 22.74 | 16.03 | 37.05 | 63.18 | 73.73 |
| PCGrad [18] | 53.37 | 74.87 | 0.3844 | 0.1611 | 22.44 | 15.79 | 37.69 | 63.53 | 73.96 |
| GradDrop [3] | | | | | OOM | | | | |
| IMTL [10] | 53.49 | 75.67 | 0.3730 | 0.1526 | 22.37 | 15.65 | 37.89 | 64.00 | 74.44 |
| GradVac [17] | 53.68 | 75.39 | 0.3811 | 0.1588 | 23.16 | 16.50 | 36.18 | 62.02 | 72.80 |
| CAGrad [9] | 53.34 | 75.42 | 0.3886 | 0.1616 | 22.61 | 16.06 | 37.07 | 63.15 | 73.89 |
| Nash-MTL [14] | 53.64 | 74.92 | 0.4014 | 0.1653 | 22.59 | 15.77 | 37.63 | 63.69 | 74.14 |
| RLW [8] | 52.25 | 74.92 | 0.3807 | 0.1561 | 22.47 | 15.73 | 37.74 | 63.75 | 74.16 |

## References


[1] Rich Caruana. Multitask learning: A knowledge-based source of inductive bias. In *International Conference on Machine Learning*, 1993.

[2] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, 2018.

[3] Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. In *Neural Information Processing Systems*, 2020.

[4] Sumanth Chennupati, Ganesh Sistu, Senthil Yogamani, and Samir A Rawashdeh. MultiNet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[5] Pengsheng Guo, Chen-Yu Lee, and Daniel Ulbricht. Learning to branch for multi-task learning. In *International Conference on Machine Learning*, 2020.

[6] Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed Chi. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. In *Neural Information Processing Systems*, 2021.

[7] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[8] Baijiong Lin, Feiyang Ye, Yu Zhang, and Ivor Tsang. Reasonable effectiveness of random weighting: A litmus test for multi-task learning. *Transactions on Machine Learning Research*, 2022.

[9] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Neural Information Processing Systems*, 2021.

[10] Liyang Liu, Yi Li, Zhanghui Kuang, Jing-Hao Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Towards impartial multi-task learning. In *International Conference on Learning Representations*, 2021.

[11] Shikun Liu, Edward Johns, and Andrew J. Davison. End-to-end multi-task learning with attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[12] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[13] 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*, 2016.

[14] Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *International Conference on Machine Learning*, 2022.

[15] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Neural Information Processing Systems*, 2018.

[16] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. Progressive layered extraction (PLE): A novel multi-task learning (MTL) model for personalized recommendations. In *ACM Conference on Recommender Systems*, 2020.

[17] Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations*, 2021.

[18] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Neural Information Processing Systems*, 2020.