Split to Learn: Gradient Split for Multi-Task Human Image Analysis Supplementary Document

A. Asymmetric Inter-task Relation

Relative performance change Consider two models with the same backbone (*i.e.*, ResNet-50-GN) with different task heads, where a single-task network is trained for task t and a multi-task network is trained to jointly minimize losses of tasks t and t'. We measure the impact of task t' on task t based on the relative accuracy change $\mathcal{A}_{t'\to t} = \frac{\text{acc}_t|\{t,t'\} - \text{acc}_t|\{t\}}{\text{acc}_t|\{t\}}$, where $\text{acc}_t|\{t,t'\}$ is the accuracy on task t of a multi-task network trained on tasks t and t', $\text{acc}_t|\{t\}$ is the accuracy of a single-task network trained on task t.

A positive value of $\mathcal{A}_{t'\to t}$ indicates that training along with task t' results in performance increase on task t, while a negative value indicates that performance decreases on task t. This inter-task relation definition is illustrated in Fig. A-1. The pairwise relative changes is summarized in Table A-1.

Effect of threshold τ Based on the relative accuracy change, we define the directive relation $t' \to t$. Concretely, if $\mathcal{A}_{t'\to t}$ is *smaller* than a threshold τ, the directive relation $t' \to t$ is defined as *negative*. We further clarify $\mathbf{m} \in \{0,1\}^{T\times T}$ defined in Eq.(2) as:

$$\mathbf{m}_{tt'} = \begin{cases} 0 & \text{if } t \neq t' \text{and relation } t' \to t \text{ is negative} \\ 1 & \text{otherwise.} \end{cases} \tag{a-1}$$

In the main paper, we used $\tau = -0.01$ for all the experiments, allowing the network to tolerate the relatively small accuracy drop of 1.00%.

To study the effect of threshold, we report the model performance for different threshold values in Table A-2: $\{\text{inf}, 0, -0.009, -0.01, -0.015, -0.025, -\text{inf}\}$. When $\tau = -\text{inf}$, GradSplit reduces to the multi-head baseline where none of the inter-task relation is negative. When $\tau = \text{inf}$, each group of filters is *only* updated by its assigned task loss, which does not fully consider the inter-task relationship during the gradient back-propagation.

As illustrated in Figure A-2, when varying τ value from $-\text{inf}$ to inf, the defined relation gradually changes. Starting from the multi-head baseline ($\tau = -\text{inf}$), overall accuracy increases as the τ increases until -0.01 (Table A-2), by masking more gradients that are from conflicting tasks (*e.g.*, Attribute $\to$ Parsing, ReID $\to$ Parsing). As τ further increases from -0.009 to inf, overall multi-task performance Δ_m decreases as the gradients become sparser. Although the gradients are sparser, $\tau = \text{inf}$ achieves slightly higher accuracy than the multi-head baseline by avoiding task conflicts. Interestingly, we observe that the multi-task performance Δ_m with $\tau = 0$ is 2.0% higher than multi-head baseline, but it is still 1.6% lower than using $\tau = -0.01$. This implies that defining negative inter-task relations with a tolerance to allow a small drop in accuracy can be beneficial, by using denser gradients during optimization.

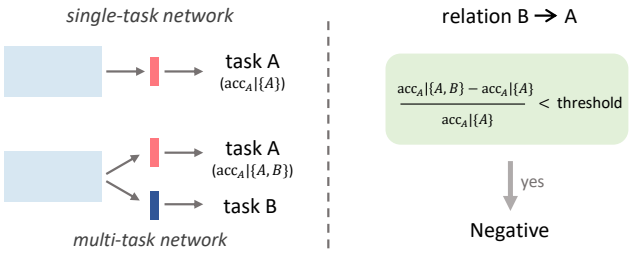

Figure A-1. Inter-task relation definition. Given a single task network for task A and a multi-task network for tasks A and B, we compare the relative performance change on task A in a validation set. If the relative change in task A accuracy, $\frac{\text{acc}_A|\{A,B\} - \text{acc}_A|\{A\}}{\text{acc}_A|\{A\}}$ is smaller than a threshold τ, the relation $B \to A$ is negative.

Table A-1. Relative performance change of each task when trained as a pair with every other task. For example, when ReID is trained with Attribute in a two-task network, Attribute performs 2.05% worse than when Attribute is trained alone. Both two-task network and single-task network use the same backbone (ResNet-50-GN). Best viewed in color.

		Relative Performance Change On			
		Attribute	ReID	Pose	Parsing
Trained With	Attribute	–	-2.16%	-1.47%	-9.87%
	ReID	-2.05%	–	-1.36%	-16.22%
	Pose	-0.77%	-0.86%	–	0.00%
	Parsing	-0.91%	-0.97%	0.11%	–

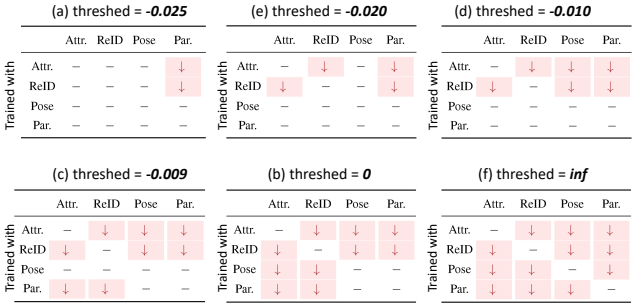

Figure A-2. Asymmetric pairwise task relations for different threshold τ. In each table, an entry (t', t) corresponds to the relation $t' \rightarrow t$. Negative relation is indicated using $\downarrow$.

Table A-2. Effect of threshold τ on GradSplit when ResNet-18-GN is used as the backbone architecture. (‡) When $\tau = -\text{inf}$, GradSplit reduces to the multi-head baseline where none of the inter-task relation is negative. When $\tau = \text{inf}$, each group of filters is *only* updated by its assigned task loss, which does not fully consider the inter-task relationship during the gradient back-propagation. (†) threshold $\tau = -0.001$ is used for all the experiment in the main paper. Overall multi-task performance Δ_m is calculated by comparing to the reference single-task networks.

Methods	Pose	Attribute	ReID	Parsing	Δ_m (↑)
	Mean (↑)	MA (↑)	mAP (↑)	mIoU (↑)	
Single-task	87.0	76.9	74.9	42.4	+0.0
Multi-head baseline ($\tau = -\text{inf}$) ‡	84.9	75.5	64.7	38.0	-7.1
$\tau = -0.025$	84.6	77.1	67.9	37.2	-6.0
$\tau = -0.015$	84.8	77.2	68.1	37.6	-5.6
$\tau = -0.010$ †	85.4	77.1	71.4	39.1	−3.5
$\tau = -0.009$	85.1	77.4	71.4	38.7	-3.7
$\tau = 0$	85.1	77.3	67.5	38.7	-5.1
$\tau = \text{inf}$	84.8	76.9	67.6	38.1	-5.6

B. Additional Experimental Results

B.1. Comparison to gradient manipulation methods

Gradient manipulation methods (*e.g.*, PCGrad and GradDrop) implicitly assume the compared task gradients to come from the *same* image [15] or domain [4]. However, in our setting, gradients of different tasks are calculated on images from different domains because each dataset has annotations only for a single task.

In Table A-3 and Table A-4, we report the results of three gradient based methods, including **PCGrad** [15], **Adversarial task disentanglement (ADV)** [12], and **GradDrop** [4].

ADV is used in ASTMT to force task gradients be statistically indistinguishable through adversarial training. While ADV brings 0.8% improvement for ASTMT on Attribute under two-task setting (Table A-3), it does not help ASTMT under three-task setting (Table A-4). Pose has much smaller loss than the remaining task. When introducing ADV, it might need carefully balancing the weights between Pose and ADV losses, which is out of scope of this work.

PCGrad [15] and GradDrop [4] directly compare task gradients, and then manipulate them to alleviate conflicts. They implicitly assume the compared task gradients to come from the *same* image or domain, which does not hold in our problem. This might the reason why the two strategies for gradient manipulation are not suitable in our setting, and achieve limited accuracy. For example, on two-task setting (Table A-3), GradDrop only achieves 35.0% mAP on ReID. One three-task setting (Table A-4), PCGrad achieves 50.8% mAP in ReID, which is 29.4% lower than our GradSplit.

Table A-3. Comparison with gradient based methods on **two** tasks: **ReID**, **Attr**. ADV denotes Adversarial task disentanglement [12]. ASTMT uses ResNet-26-TBN as backbone for a fair comparison.

Methods	Backbone	Attribute	ReID	
		MA (↑)	Rank-1 (↑)	mAP (↑)
ASTMT [12]	R26-TBN	76.6	89.3	73.4
ASTMT + ADV [12]	R26-TBN	77.4	89.4	73.0
GradDrop [4]	R50-GN	74.5	60.2	35.0
PCGrad [15]	R50-GN	74.8	91.6	74.6
Multi-head baseline	R50-GN	76.4	90.8	76.9
GradSplit		78.0	92.1	79.9

Table A-4. Comparison on **three** tasks: **ReID**, **Attribute**, and **Pose**. Overall multi-task performance Δ_m is calculated by comparing to the reference single-task networks (ResNet-50-GN)

Methods	Backbone	Attribute	ReID	Pose	Δ_m (↑)	#Param (M)
		MA (↑)	mAP (↑)	Mean (↑)		
Single-task	R50-GN	78.0	81.1	88.2	+0.0	85
PCGrad [15]	Res.50-GN	50.0	50.8	16.2	-51.6	38
ASTMT [12]	R50-TBN	79.5	59.0	86.4	-9.1	48
ASTMT + ADV	R50-TBN	78.2	74.1	63.1	-12.3	48
GradNorm [3]	R50-GN	74.0	54.5	85.1	-13.8	38
MTAN [9]		77.4	50.0	85.5	-14.0	38
Multi-head	R50-GN	75.9	76.5	86.3	-3.5	38
GradSplit		77.6	80.2	86.3	**-1.3**	38

B.2. Comparison to task balancing methods

In our problem, the magnitude of pose loss is much smaller than the magnitudes of other task losses. To this end, we include two task balancing methods in the experiment and report results in Table A-4. **GradNorm [3]** aims to balance task losses by stimulating the task-specific gradients to be of similar magnitude. However, GradNorm failed to handle this imbalance effectively and achieved low overall accuracy. **MGDA [14]** seeks to find Pareto optimal solutions. We observe that it still cannot address imbalanced losses, and produces undesirable performance on three-task setting (*e.g.*, 16.0% in Pose).

B.3. Result on three-task setting

In Table A-5, we report the results under three-task setting (Parsing, Pose, and Attr). We also observe that GradSplit achieves +4.8% higher overall multi-task performance than multi-head baseline.

Table A-5. Comparison on **three** tasks: **ReID**, **Attr**, and **Parsing**.

Methods	Backbone	Attribute	ReID	Parsing	Δ_m	#Param
		MA (↑)	mAP (↑)	mIoU (↑)	(↑)	(M) ↓
Single-task	R50-GN	78.0	81.1	45.6	+0.0	89
	R18-GN	76.9	74.9	42.4	–	47
Multi-head	R50-GN	76.6	79.0	38.8	-6.4	44
GradSplit		77.2	80.4	43.5	**-1.6**	44

Table A-6. Results on **synthetic** dataset: Digit classification, CIFAR image recognition and Digit segmentation. We report the two-task performance to show pair-wise task relation. We compare the multi-head baseline and GradSplit under the three-task setting.

Methods	Backbone	MNIST	CIFAR	Parsing
		Acc. (↑)	Acc. (↑)	mIoU (↑)
Single-task	R18-GN	91.7	75.8	78.3
Two-task	R18-GN	84.5	67.8	-
		92.9	-	77.6
		-	85.2	47.1
Multi-head	R18-GN	85.0	73.0	73.6
GradSplit		88.8	76.4	73.9

C. Understand GradSplit with Synthetic Dataset

We use MNIST and CIFAR-10 to create a three-task setting: CIFAR classification, Digit classification, and Digit segmentation. An example image (left) and its ground truth segmentation (right) are shown in Fig. A-3. We generate an image by mixing randomly selected images from MNIST and CIFAR-10. Specifically, we overlay the digit on the random position of the image from CIFAR-10. There are conflicts among these three tasks. Segmentation task discriminates the foreground region (MNIST) and background region (CIFAR), so it benefits the classification tasks. However, the opposite is not true. For example, MNIST classification needs to distinguish different digits, while segmentation needs digit-agnostic features.

GradSplit is beneficial, because it improves model performance by reducing gradient conflicts during training, while still exploiting task-specific features across tasks. As shown in Table A-6, GradSplit consistently outperforms the multi-head baseline on all three tasks.

Figure A-3. Illustration of synthetic dataset. It has three tasks: Digit classification, CIFAR-10 image recognition, and Digit segmentation. Asymmetric relations are apparent under this setting. For example, segmentation features help classification tasks, but the opposite is not true.

D. Implementation Details

D.1. Multi-head baseline

Data pre-processing and augmentation To maximize accuracy while maintaining efficient computation, we allowed different sizes of inputs for tasks. Specifically, we resized the input image to 288×288 for Parsing, 256×256 for pose estimation, and 256×128 for person re-identification and attribute recognition. Random flip and RandomErasing are used for Attribute and ReID; Multi-scale crop and random flip are used for Parsing; Multi-scale crop, random rotation($\pm$40 degrees) and flip are used for Pose.

Mini-batch construction ReID uses Random-Identity Sampler [7] to sample 64 images from 4 identities per mini-batch. Attribute, Parsing, and Pose use Random-Sampler to sample 64 images for each mini-batch.

Loss Pose uses MSE loss, Parsing uses pixel-wise cross entropy loss, Attribute uses binary cross-entropy loss for each attribute, ReID uses triplet loss and cross-entropy loss. We adopt round-robin batch-level update regime [11] for optimization. One multi-task iteration consists of a sequence of each task batch forwarding and parameter updating. Namely, for each iteration, we only train network on samples from one task. With this strategy, we tried using large weights for Pose (its loss is relative small than the other tasks) but empirically observed its accuracy slightly changed. Thus, we used uniform weights in all the experiments for simplicity.

D.2. Comparing methods

Cross-stitch Network [13], NDDR [5] Following the paper, we first train task-specific networks separately and then finetune the whole network, including the interaction modules, to minimize the joint task loss. We used $\alpha = 0.9$ and $\beta = 1/(N-1)$ (N is the number of tasks). The interaction modules are inserted from layer 1 to layer 4 of the ResNet. **ASTMT [12]**, **MGDA [14]**, **RCM [8]**, **MTAN [9]** and **SFG [2]** are implemented based on the official codes.

PCGrad [15] PCGrad is implemented based on the official code and applied to the last layer of the shared backbone.

GradDrop [4] GradDrop layer is applied to the *last common feature maps* which are the inputs to the task-specific heads. GradDrop requires the size of the feature maps for all tasks to be the same so that the feature map values across tasks are element-wise comparable. Since we used different sizes of inputs for tasks, we tested GradDrop on the two-task setting, *i.e*., person re-identification and attribute recognition, where inputs are resized to the same size.

GradNorm [3] We applied GradNorm to the last layer of the common feature extractor. We normalized gradients before calculating the gradient norm loss.

E. Potential Negative Social Impact

Since our work is related to human recognition, there is a potential risk that our work can be utilized for unlawful surveillance. To mitigate the unintended use of our work, the code and model with downstream applications will be accompanied with the precautions to highlight this risk. Our work relied on the datasets consisting of human images, which could be subject to privacy risks. We used publicly released datasets and followed the protocol used in prior works for evaluation (*i.e.*, MPII [1], PA-100k [10], LIP [6], Market-1501 [16]).

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[2] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 1385–1394, 2019.

[3] 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*, pages 794–803. PMLR, 2018.

[4] 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 *NeurIPS*, 2020.

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3205–3214, 2019.

[6] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017.

[7] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[8] 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 *Proceedings of the European conference on computer vision (ECCV)*, 2020.

[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.

[10] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017.

[11] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

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

[13] 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.

[14] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, pages 527–538, 2018.

[15] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *NeurIPS*, 2020.

[16] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.