Task Meta-Transfer from Limited Parallel Labels

Supplementary Materials

Anonymous Author(s)
Affiliation
Address
email

A. Algorithm

In this section, we provide our pseudo-code together to further illustrate our method, Task Meta-Transfer.

Algorithm 1 *Task Meta-Transfer*

Require: Network parameters Θ, $\theta_{\mathcal{A}}$ and $\theta_{\mathcal{P}}$, gradient projection ϕ, learning rate α, update frequency k, total training steps T

```
while step < T do
    Algorithm 3
    if step%k = 0 then
        Algorithm 2
    end if
end while
```

Algorithm 2 *Meta-Learning from Parallel Labels*

Require: Network parameters Θ, $\theta_{\mathcal{A}}$ and $\theta_{\mathcal{P}}$, gradient projection ϕ, learning rate α

$(x, y^{\mathcal{A}}, y^{\mathcal{P}}) \leftarrow \text{RandomSample}(\mathcal{X}, \mathcal{Y}^{\mathcal{A}}, \mathcal{Y}^{\mathcal{P}})$ ▷ Sample a mini-batch

$o = f_{\Theta}(x)$ ▷ Calculate BackBone's activation o

$\mathcal{L}_{\mathcal{A}} = \mathcal{L}(f_{\theta_{\mathcal{A}}}(o), y^{\mathcal{A}})$ ▷ Calculate the auxiliary loss

$g_{\mathcal{A}} = \nabla_o \mathcal{L}_{\mathcal{A}}$ ▷ Calculate the gradients w.r.t activation o

$g'_{\mathcal{A}} = f_{\phi}(g_{\mathcal{A}})$ ▷ Project the auxiliary gradients

$\Theta^{+} = \Theta - \alpha g'_{\mathcal{A}} \frac{\partial o}{\partial \Theta}$ ▷ Retain computational graph and compute Θ^{+}

$\mathcal{L}_{\mathcal{P}} = \mathcal{L}(f_{\theta_{\mathcal{P}}}(f_{\Theta^{+}(\phi)}(x)), y^{\mathcal{P}})$ ▷ Forward pass x to Θ^{+} to compute the primary loss

$\phi \leftarrow \phi - \beta \nabla_{\Theta^{+}} \mathcal{L}_{\mathcal{P}} \nabla_{\phi} \Theta^{+}$ ▷ Update f_{ϕ}

$\Theta \leftarrow \Theta - \alpha \frac{\partial \mathcal{L}_{\mathcal{P}}}{\partial \Theta}$ ▷ Update Θ

$\theta_{\mathcal{P}} \leftarrow \theta_{\mathcal{P}} - \alpha \frac{\partial \mathcal{L}_{\mathcal{P}}}{\partial \theta_{\mathcal{P}}}$ ▷ Update $\theta_{\mathcal{P}}$

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Algorithm 3 *Learning from Auxiliary Data*

Require: Network parameters Θ and $\theta_{\mathcal{A}}$, gradient projection ϕ, learning rate α

$(x, y^{\mathcal{A}}) \leftarrow$ RandomSample $(\mathcal{X}, \mathcal{Y}^{\mathcal{A}})$ ▷ Sample a mini-batch

$o = f_{\Theta}(x)$ ▷ Calculate BackBone's activation o

$\mathcal{L}_{\mathcal{A}} = \mathcal{L}(f_{\theta_{\mathcal{A}}}(o), y^{\mathcal{A}})$ ▷ Calculate the auxiliary loss

$g_{\mathcal{A}} = \nabla_o \mathcal{L}_{\mathcal{A}}$ ▷ Calculate the gradients w.r.t activation o

$g'_{\mathcal{A}} = f_{\phi}(g_{\mathcal{A}})$ ▷ Project the auxiliary gradients

$\Theta \leftarrow \Theta - \alpha g'_{\mathcal{A}} \frac{\partial o}{\partial \Theta}$ ▷ Update Θ

$\theta_{\mathcal{A}} \leftarrow \theta_{\mathcal{A}} - \alpha \frac{\partial \mathcal{L}_{\mathcal{A}}}{\partial \theta_{\mathcal{A}}}$ ▷ Update $\theta_{\mathcal{A}}$

B. Datasets

NYUv2 [7] We use the official training split of 795 images for training, and partition the original testing split of 645 images into two subsets of 327 images to obtain a validation set and a test set with identical sizes (using even indices for validation and odd indices for testing). We simulate a scenario where limited labels for the primary task are available by assuming that the set of parallel examples $\mathcal{X}$ is a small subset of the entire training set (we experiment with different sizes of set $\mathcal{X}$), while the auxiliary training set $\mathcal{X}^{\mathcal{A}}$ consists of all 795 examples. In order to conduct exhaustive ablation studies and to collect statistics from multiple training runs for each model, we reduce the computational cost of training by resizing the images to size 288×384.

CityScapes [1] is a high resolution dataset with parallel labels, mainly used for semantic urban scene understanding. The benchmark includes a training split of 2975 images with pixel-level annotations which we use for training. We partition the official validation split of 500 images into a test set and a validation set of 250 images each. Our experiments include auxiliary-primary task-pairs obtained from the tasks of inverse depth estimation, 7-class segmentation [6] and 19-class segmentation. We do not consider the auxiliary-primary pair of 7-class segmentation and 19-class segmentation as these tasks are closely related and overlap in label space, whereas we are interested in scenarios where the tasks are quite distant. We reduce the input image size to 128×256 for speeding up our training.

C. Network Architectures

SplitVGG has an encoder which is topologically identical to VGG16 [8], and a decoder having the same structure but with transposed convolutional layers performing upsampling in lieu of the downsampling convolutional layers of the encoder. Both the encoder and the decoder have 5 blocks, each block including either 2 or 3 stacks of Conv-BN-Relu layers. The official VGG-16 has $[64, 128, 256, 512, 512]$ channels at each block. We implement our network using the same topology but using half the number of channels at each block, i.e., $[32, 64, 128, 256, 256]$. The upsampling decoder in SplitVGG has 5 blocks with channels of $[256, 256, 128, 64, 32]$. All convolutional layers are implemented using kernels of 3×3, stride of 1×1, padding of 1×1 and no bias. Transposed convolutional layers have kernels with size of 3×3, stride of 2×2, padding of 1×1 and no bias. The total number of parameters in SplitVGG is 7.34M.

SplitRes has a 5-block downsampling encoder which is topologically identical to ResNet18 [3], and a corresponding upsampling decoder with 5 blocks. The official ResNet18 has $[64, 64, 128, 256, 512]$ channels at each block. We adopt the same architecture but reduce the number of channels at each block to $[32, 32, 64, 128, 256]$. The upsampling decoder has 5 blocks similar to SplitVGG with number of channels of $[128, 128, 64, 32, 32]$. The total number of parameters in SplitRes is 5.62M.

The smaller number of parameters in SplitRes leads to overall inferior performances across all methods comparing to SplitVGG in the main paper. We also note that our networks are much smaller than the one in Liu et al. [6] with 24.98M parameters. Thus, the baseline performances reported in our paper have larger errors compared to those in Liu et al. [6].

D. Training Details

For TMT, multi-task learning and the pretraining of transfer learning, we use SGD with a learning rate initially set to 0.1 on all parameters of both backbone and primary network and reduce it by half in the middle of training. On NYUv2 we train models for 40K steps with a batch size of 2, while on CityScapes we train for 44K steps with a batch size of 8.

For the fine-tuning stage of transfer learning on the primary tasks of semantic segmentation and surface normal estimation, we use SGD with initial learning rate of 0.005 for the backbone f_Θ and 0.05 for the newly attached primary network $f_{\theta_\mathcal{P}}$. Instead, for depth estimation, we use 0.01 and 0.1, respectively. These differences in learning rates among the tasks reflect our best attempt at optimizing performance for this baseline. The fine-tuning has 8K steps for both NYUv2 and CityScapes, with batch size of 2 and 8, respectively. We decrease the learning rate by half at 4K steps for both datasets.

For training the gradient projection, we use Adam optimizer with a learning rate of 0.001 and weight decay set to 10^{-5} on NYUv2, while on CityScapes the learning rate is 0.01 and the weight decay is 10^{-5}. The learning rate is decreased by half in the middle of training.

Early Stopping For each training, we store a checkpoint every 200 steps and select 10 consecutive checkpoints with the lowest average validation error. We report test performance by averaging the test errors for those 10 consecutive models. By doing this, we eliminate the chance of choosing a model in the unstable training.

E. Training Errors for Different Ranks of Projection Model

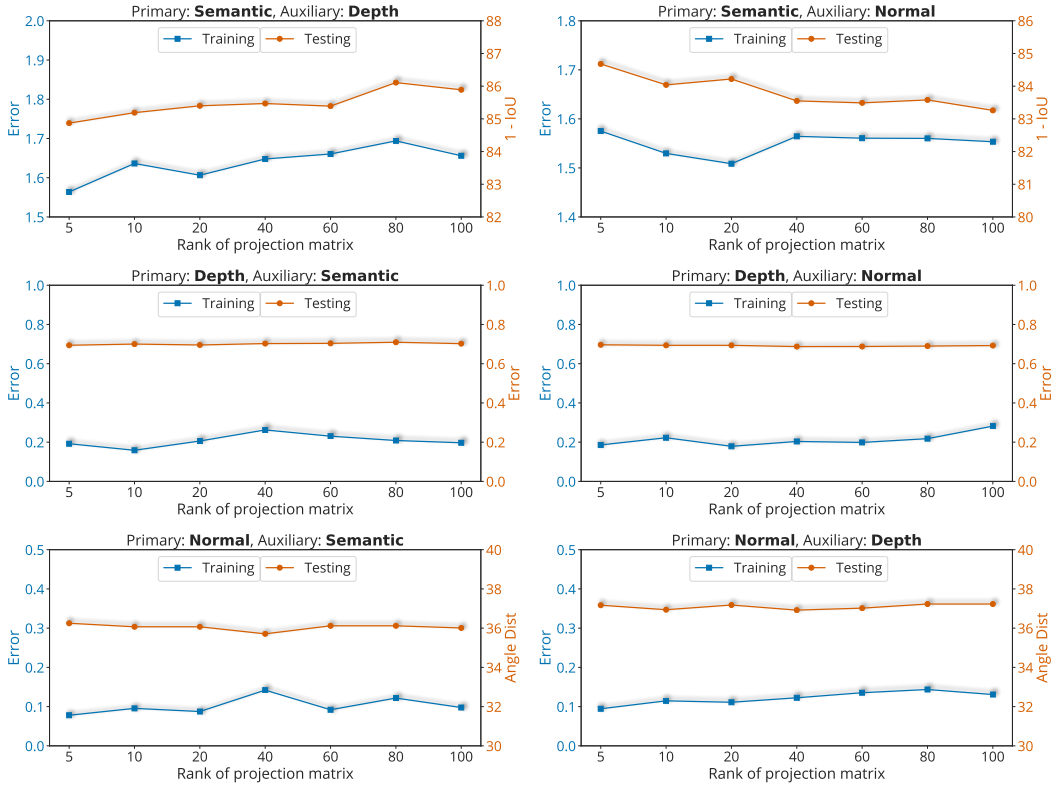

Figure 1: Training error and testing performance for varying the rank r of the matrix defining the gradient transformation f_ϕ on NYUv2 using SplitVGG-3.

Increasing rank r of gradient projection f_ϕ yields substantial improvements on "semantic using normal". While in "semantic using depth" it leads to degradation in performance. In both cases, we find that better models are trained at lower training errors. This implies that in "semantic using depth", an over-parameterized projection f_ϕ makes it difficult to optimized during meta-learning

steps. In contrast, in "semantic using normal" better performance can be obtained through a higher-parameterized projection f_ϕ that yields lower training error. For the other 4 cases involving primary tasks of "depth" and "normal", the accuracy of TMT remains relatively unperturbed with different choices of rank r in projection f_ϕ.

F. Statistical Variance and Hyper-parameters

Multi-Task Learning (MTL) is the joint training of the entire model with respect to both the primary and the auxiliary tasks. We search a hyper-parameter k (see in Algorithm 1) from validation set that controls how often we alternate training between primary task and auxiliary task. As shown in the main text, this method gives overall better results comparing to other advanced MTL methods, that are designed for cases where labels are available for the full dataset. Whereas in our study, only a small subset examples have labels in both primary and auxiliary tasks. *Task Uncertainty Weighting* by [4] and *Cosine Similarity Weighting* by [2] both have their mechanisms to weight the tasks and are designed to avoid extensive hyper-parameters search (Though in the results of main text, we add the same alternating mechanism to get the best results). Our studies indicate that those mechanisms probably do not work well in such imbalanced datasets, comparing to directly search for the fixed weighting hyper-parameters k.

The same hyper-parameter k is also used in the two meta-methods: Meta Weighting [5] and our method *TMT*. *Task Meta-Transfer (TMT)* has the same hyper-parameter k, together with rank r of gradient projection, which controls the complexity of projection models ϕ.

We report the hyper-parameters k and r chosen from the cross validation, and the statistical variance in Table 1. Note that the robustness study of the experiments in main text shows a wide range of r works well in our method.

Table 1: Statistical variance and learned hyper-parameters of different models (average of 3 runs) on *NYUv2* dataset using different pairs of primary and auxiliary tasks with two distinct architectures (SplitVGG-3 and SplitRes-3).

	Tasks		*Transfer Learning*	*weighted MTL*		*[4]*		*Du et al. [2]*		*Lin et al. [5]*		*TMT (ours)*		
	Primary	*Auxiliary*	var	var	k	var	k	var	k	var	k	var	k	r
SplitVGG-3	Semantic (IoU ↑)	Depth	0.71	0.25	9	0.55	7	0.28	9	0.21	10	0.35	5	5
		Normal	0.63	0.95	9	0.14	7	0.39	9	0.32	6	1.03	8	100
	Depth ($10^2 \times$AbsError↓)	Semantic	0.53	0.38	4	0.35	4	0.32	5	0.32	9	0.24	4	5
		Normal	0.32	0.09	7	0.76	9	0.30	9	0.94	7	0.36	5	40
	Normal (AngleDist↓)	Normal	0.26	0.07	2	0.16	5	0.38	4	0.11	5	0.20	2	40
		Depth	0.06	0.14	5	0.29	9	0.21	8	0.15	9	0.13	9	40
SplitRes-3	Semantic (IoU ↑)	Depth	0.38	0.09	9	0.09	9	0.14	10	0.19	10	0.35	10	60
		Normal	0.86	0.67	9	0.67	9	0.13	7	0.80	7	0.63	9	60
	Depth ($10^2 \times$AbsError↓)	Semantic	0.02	0.06	3	0.98	5	0.15	6	0.17	7	0.99	3	80
		Normal	0.37	0.07	10	1.60	7	0.33	8	0.13	8	0.47	9	60
	Normal (AngleDist↓)	Semantic	0.06	0.10	2	0.08	10	0.08	4	0.30	6	0.14	5	40
		Depth	0.17	0.05	10	0.19	9	0.18	8	0.26	9	0.04	9	40

References

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] Y. Du, W. M. Czarnecki, S. M. Jayakumar, R. Pascanu, and B. Lakshminarayanan. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*, 2018.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[4] A. Kendall, Y. Gal, and R. 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 (CVPR)*, 2018.

[5] X. Lin, H. Baweja, G. Kantor, and D. Held. Adaptive auxiliary task weighting for reinforcement learning. In *Advances in Neural Information Processing Systems 32*, 2019.

[6] S. Liu, E. Johns, and A. 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.

[7] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Response to reviewer 1

Weakness (1) Missing related work of Lin et al. [2019] has been added to paper in Line 99. (2) We remove the claim of adaptation of self/un-supervised learning, and we didn't talk about the reduction of computational cost as number of auxiliary tasks increases. (3) We have added an advanced multi-task method Kendall et al. [2018] and meta adaptive weight Lin et al. [2019] in our Section 4.3 and Table 1 and Table 2 for comparison.

Relation to prior work (1) We add reference of open-set and partial-set domain adaptation in Line 47 and 48. (2) We take the advise and adapt Lin et al. [2019] to our setting and show results in experiments section.

Additional feed back (1) In Algorithm 3 (in supplymentary) each auxiliary task still requires gradient backpropagation, only for the purpose of updating the auxiliary head network $\theta_{\mathcal{A}}$, but not the shared backbone Θ. This allows for obtaining meaningful auxiliary gradients for transformation. (2) Comparison to Lin et al. [2019] is added in experiment section. (3) Statistical variance are now added to supplementary. (4) Hyper-parameters used in experiments are now added to supplementary. (5) This is exactly what we do in the paper. We do not set the auxiliary set the same size of primary target set. The primary set is only 10% of auxiliary set in most parts of experiments. And we vary the ratio to 20%, 33% 50% in Section 4.5.

Response to reviewer 2

Weakness We add Figure 2 in main paper to further illustrate our approach. We study two dataset, both with tasks in semantic segmentation, depth estimation and surface normal estimation. In our experiment (Table 1, Table 2), we pick one of the three task as primary, one of the rest two task as auxiliary, this leads to six primary-auxliary pairs studied throughout the paper (The only exception is Table 3, where we pick one task as primary and both two other tasks as auxiliary).

Response to reviewer 3

Weakness (1) We add in Line 37, the motivation is that we can harness information from label spaces that are easier to annotate (e.g., depth maps collected automatically via laser scanners) to improve performance on tasks where labels are costly to obtain (e.g., semantic segmentation maps, which require each pixel to be manually annotated by a human observer). (2) Our goal is to carry out empirical study in the paper. (3) Multi-label learning is not strongly related to our work, and we do not add them in the main paper due to the lacking of space.

Response to reviewer 4

Weakness (1) We add an advanced multi-task method Kendall et al. [2018] and meta adaptive weight Lin et al. [2019] in our Section 4.3 and Table 1 and Table 2 for comparison. (2) We provide the training error for different ranks of projection model in supplementary. (3) In Figure 3, we show that in many cases, increasing the complexity of projection model will not lead to much better performance. The linear projection study in our paper should be sufficient enough. (4) Our intention is to show the proof of concept, instead of getting SoTA. Due to the demanding memory requirement and training time for the meta learning step, we use a different much smaller network as base model, comparing to those SoTA methods.

Correctness (3) In Line 178, the gradient passed into Θ is not $g_{\mathcal{A}}$ (, which may cause task interference) but the transformed $f_\phi(g_{\mathcal{A}})$. And ϕ is optimized for primary task. (4) We add in Line 200 to further describe the segmentation split and surface normal in our study of NYUv2.

References

A. Kendall, Y. Gal, and R. 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 (CVPR)*, 2018.

X. Lin, H. Baweja, G. Kantor, and D. Held. Adaptive auxiliary task weighting for reinforcement learning. In *Advances in Neural Information Processing Systems 32*, 2019.

View Reviews

Paper ID
6284

Paper Title
Task Meta-Transfer from Limited Parallel Labels

Reviewer #1

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
In this work, the authors proposed to adapt the gradients of auxiliary tasks to improve the performance in the primary task, where the adaptation function in the form of a linear matrix is learned by gradients of gradients. Basically, this paper follows the line of multitask learning, except that the authors claimed to focus on the primary task only.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
1) The idea of adapting the gradients directly to the primary task via a linear projection is novel.
2) The number of parallel labels required for both tasks, i.e., the auxiliary task and the primary task, could be small, which is beneficial when the primary task is challenging to annotate.
3) The paper is well written and easy to follow.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
1) There are some missing works, e.g., [1], in the literature review, which should also be compared empirically to validate the effectiveness of the proposed algorithm.
2) A lot of aspects that claimed have not been fully investigated and verified, including the accommodation of self-supervised/unsupervised auxiliary tasks and the reduction of computational costs when the number of auxiliary tasks increases.
3) The improvement of the proposed is minor, especially considering that some state-of-the-art baselines have not been included.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
The paper is technically right.

5. Clarity: Is the paper well written?
Overall, this paper is clearly written, though there are minor grammatical mistakes:
- Line 155: use -> used
- Line 187&189: no closing marks
- Line 266: it's -> it is
- Line 269: i.e -> i.e.,

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
1) There are some misappropriate claims about prior works, which is listed as below.
- Domain adaptation is not limited to the case where the output spaces of the two domains are the same. Please kindly refer to related works on partial-set and open-set domain adaptation.
2) There are some missing related works in the literature of multitask learning.
- Since the intuition behind this work is to learn the relationship between the primary task and the auxiliary task, it is expected to comprehensively survey the related works on multitask learning which learn the relationship.
- As mentioned in the weakness, the most recent work [1] by adapting the auxiliary task and also meta-learning the

weights for the auxiliary task is missing. More importantly, this work is also specifically designed to improve the performance for the primary task only, which is exactly the same with this work. Though this work is empirically validated on reinforcement learning problems, it is completely applicable on supervised learning problems studied in this work.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

8. Additional feedback, comments, suggestions for improvement and questions for the authors:
1) The authors claimed that one of the advantages of this work is its avoidance of task interference, as the auxiliary tasks do not consume the learning capacity. Unfortunately, as shown in Algorithm 3 in the supplementary, each branch for each auxiliary task still requires gradient backpropagation. Therefore, I suppose that the claim made by the authors is problematic and requires extensive empirical validation of the computation cost/performance as the number of auxiliary tasks increases. Moreover, by comparing Table 1 and Table 3, the proposed TMT indeed suffers from task interference -- more auxiliary tasks do not contribute much better performances.
2) The effectiveness of the adaptation matrix requires further explanation and consolidation.
- It is not intuitive that the performance stay all consistent as the rank r varies considerably from 5 to 100. Therefore, I am even curious about the performance of a random adaptation matrix or a linear adaptation scalar (like a weight). It learning a linear scalar does hold, then this work is actually the same/similar to [1], which simply learns a weight for the auxiliary task.
3) The statistical variance should be reported, as the performance improvement is marginal compared to the baselines.
4) The learned hyperparameters via cross-validation, e.g., k, should be reported for all algorithms.
5) Could you report the results that conducting gradient descent using X more times than X_A? The performance is not necessarily bad, to my knowledge.

[1] Lin, X., Baweja, H., Kantor, G., & Held, D. (2019). Adaptive Auxiliary Task Weighting for Reinforcement Learning. In Advances in Neural Information Processing Systems (pp. 4772-4783).

9. Please provide an "overall score" for this submission.
5: Marginally below the acceptance threshold.

10. Please provide a "confidence score" for your assessment of this submission.
4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes

Reviewer #2

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
The author proposes a meta-learning algorithm using information from “parallel labels” between the primary task and the auxiliary task. The experiment shows the proposed algorithm has an advantage over naïve transfer learning, multi-task learning, and other related work. The authors also claim the proposed algorithm has no requirement for the degree of overlapping.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.

- I think the major contribution of this paper is the learnable gradient projection function f_phi, which has the potential to be used in other areas.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
- The abstract claims the algorithm has no requirement on the degree of overlapping between labels. However, as the major contribution of the paper is to use such overlapping information, I wonder if the author can find some measurement on the degree of label overlapping (i.e. the difference between segmentation label and normal label, since from figure 3 it seems the normal label contributes to the semantic labeling by a large margin).

- In the normal estimation experiment in figure 2, the proposed method does not perform better than multi-task learning and transfer learning. Please provide some analysis on this.

- Figure 1 in the supplementary materials should be in the main text as it defines the workflow. The text in figure 1 does not involve how we can update f_phi, which should be the major contribution. At this stage, the caption in figure 1 is misleading since the reader might get confused about why f_phi can be updated as it is not in the forward pass.
- The authors mentioned the method can be used other than the computer vision area. However, the authors did not include experiments to support this claim. Furthermore, the authors only report experiments on scene understanding. What if this algorithm is implemented on the bounding box level (i.e. multi-task re-ID [1])?

Minor comments.
- Please shows more details in figure 1: what is the primary label and what is the auxiliary label? From this figure, it is hard for us to tell the difference. The illustration in Section 3.2 does not have a clear description, either. (From the experiment I guess the auxiliary label should be the depth image).

- Please make the reference format consistent. In reference 3, it shows ‘international conference on international conference on machine learning’. In reference 41, it shows ‘in in computer vision-ECCV’.

- Line 187, a ‘)’ is missing

[1] Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C., & Yang, Y. (2019). Improving person re-identification by attribute and identity learning. Pattern Recognition, 95, 151-161.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
yes.

5. Clarity: Is the paper well written?
mostly yes.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
yes.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

9. Please provide an "overall score" for this submission.
6: Marginally above the acceptance threshold.

10. Please provide a "confidence score" for your assessment of this submission.
2: You are willing to defend your assessment, but it is quite likely that you did not understand central parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes

Reviewer #3

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.
This paper proposes a meta-learning method that uses auxiliary tasks to improve the model on the primary task, assuming the existence of a small training set with parallel labels. Experiments are conducted on two image analysis benchmarks involving multiple tasks by comparing with naïve transfer learning, multi-task learning and some prior related work.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
The proposed approach is novel and interesting.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
1. Motivation for the problem and approach in the introduction is poor. Transfer learning mainly tackles the domain divergence between source and target domains. The studied setting is very different from transfer learning. It is better to be motivated from possible real world demands in such settings. An example of the real-world scenarios that fits the problem setting can be helpful.

2. Strong assumption: It states “Our only underlying assumption is that the tasks are loosely related such that gradients computed for the auxiliary task may be “transformed” into gradients useful to optimize the model with respect to the primary task. “. But how can the loosely relatedness guarantee a useful gradient transformation? Some theoretical proof can greatly strengthen the work.

3. Missing related works on multi-label learning with incomplete labels.

4. The assumption of the existence of a hold-out validation set does not make sense in the label-rare (primary task) training setting.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
Seem to be correct.

5. Clarity: Is the paper well written?
The motivation of the paper and approach in the introduction can be improved.

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
Relationship with multi-task learning is not well positioned. It states the multi-task learning must address all tasks simultaneously. This certainly is not true. Multi-task learning can separate auxiliary tasks from the main task as well.

Multi-label learning with incomplete labels will be highly related with the proposed problem. Related works on this topic are entirely ignored.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

9. Please provide an "overall score" for this submission.
5: Marginally below the acceptance threshold.

10. Please provide a "confidence score" for your assessment of this submission.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes

Reviewer #4

Questions

1. Summary and contributions: Briefly summarize the paper and its contributions.

The method introduces a meta-learning algorithm to transform gradient information from auxiliary tasks into gradients for a primary task in order to improve performance on a primary task. The focus of the method is on only requiring a small training set with parallel labels in order to learn a primary task.

2. Strengths: Describe the strengths of the work. Typical criteria include: soundness of the claims (theoretical grounding, empirical evaluation), significance and novelty of the contribution, and relevance to the NeurIPS community.
The method is interesting and only requires a small parallel label training set.
The projection of the gradients is a unique idea worth attempting. Results are shown with primary task labels being less than 10% of auxiliary examples. The paper is clear and well written.

3. Weaknesses: Explain the limitations of this work along the same axes as above.
There is no comparison between other methods that the method is close to such as [5, 14, 15].

The paper doesn't discuss overfitting with the primary labels. What if the primary labels are all similar scenarios and aren't a diverse distribution of samples. Would that change learning the parameterization from auxiliary gradient to primary gradient?

The paper mentions that their experiments show that simple linear projections are sufficient in Line 199. However, this isn't mentioned anywhere else or shown in any ablation studies.

The results on NYUv2 are very low mIoU. How meaningful is this? Many multi-task methods have mIoU of 33-38 [A,B]. The surface normal mean angle is also very high compared to SoTA which is as low as 17-20 [C,D,E]. These results aren't very compelling given they are much lower than current best methods.

The improvement over MTL is very limited. Most of the listed improvements are less than 0.5 more than MTL which doesn't feel like a substantial improvement or difference over standard MTL with hard parameter sharing, let alone more advanced hard/soft parameter sharing methods. Along with this, 18 is the only comparison shown compared to many others out there. There is very limited comparison to other state of the art multitask and auxiliary loss learning papers, many the paper even cites.

The paper introduces and interesting concept but the experimental evaluation is lacking and the results are not that much better than standard MTL.

A] A. Mousavian, H. Pirsiavash, and J. Koseck ˇ a, "Joint semantic segmentation and ´ depth estimation with deep convolutional networks," in 2016 Fourth International Conference on 3D Vision (3DV), IEEE, 2016, pp. 611–619.
B] Y. Gao, J. Ma, M. Zhao, W. Liu, and A. L. Yuille, "Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
C] Hickson, Steven, et al. "Floors are flat: Leveraging semantics for real-time surface normal prediction." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019.

D] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 283–291, 2018.
E] P. Wang, X. Shen, B. Russell, S. Cohen, B. Price, and A. L. Yuille. Surge: Surface regularized geometry estimation from a single image. In Advances in Neural Information Processing Systems, pages 172–180, 2016.

4. Correctness: Are the claims and method correct? Is the empirical methodology correct?
Line 42 specifies strong performance gains, however, the gains shown in Tables 1-3 don't show this.
Line 45-47 claims the technique can be applied involving multiple auxiliary tasks, yet this isn't shown in the paper. Only 2 auxiliary tasks are shown at a time.

The paper claims the approach is not specific to the computer vision domain but this isn't shown anywhere.

Line 89-91: Nowhere is it proven that the method is immune to task interference.

The NYUv2 semantic split is not mentioned (13, 20, 40, etc).
The type of surface normals for NYUv2 are not shown or cited appropriately. There are several different surface normals produced for this dataset. Which did the authors use? (Ladicky, Eigen, Hickson, etc. all have different surface normals produced for this dataset)

5. Clarity: Is the paper well written?
Yes

6. Relation to prior work: Is it clearly discussed how this work differs from previous contributions?
Yes though not compared against.

7. Reproducibility: Are there enough details to reproduce the major results of this work?
Yes

9. Please provide an "overall score" for this submission.
3: A clear reject.

10. Please provide a "confidence score" for your assessment of this submission.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

11. Have the authors adequately addressed the broader impact of their work, including potential negative ethical and societal implications of their work?
Yes