Active Learning with Task Consistency and Diversity in Multi-Task Networks

Aral Hekimoglu
Technical University Munich
Munich, Germany
aral.hekimoglu@tum.de

Michael Schmidt
BMW Group
Munich, Germany
michael.se.schmidt@bmw.de

Alvaro Marcos-Ramiro
BMW Group
Munich, Germany
alvaro.marcos-ramiro@bmw.de

Abstract

Multi-task networks demonstrate state-of-the-art performance across various vision tasks. However, their performance relies on large-scale annotated datasets, demanding extensive labeling efforts, especially as the number of tasks to label increases. In this paper, we introduce an active learning framework consisting of a data selection strategy that identifies the most informative unlabeled samples and a training strategy that ensures balanced training across multiple tasks. Our selection strategy leverages the inconsistency between initial and refined task predictions generated by recent two-stage multi-task networks. We further enhance our selection by incorporating task-specific sample diversity through a novel feature extraction mechanism. Our method captures task features for all tasks and distills them into a unified representation, which is used to curate a training set encapsulating diverse task-specific scenarios. In our training strategy, we introduce a sample-specific loss weighting mechanism based on the individual task selection scores. This facilitates the individual prioritization of samples for each task, effectively simulating the sample ordering process inherent in single-task active learning. Extensive experimentation on the PASCAL and NYUD-v2 datasets demonstrates that our approach outperforms existing state-of-the-art methods. Our approach reaches the loss of the network trained with all the available data using only 50% of the data, corresponding to 10% fewer labels compared to the state-of-the-art selection strategy. Our code is available at https://github.com/aralhekimoglu/mtal.

1. Introduction

Multi-task networks have demonstrated state-of-the-art (SOTA) performance across a range of vision tasks, including semantic segmentation [3, 33] and monocular depth estimation [16, 22]. The performance of these multi-task networks depends on the availability of labels for multiple tasks for each sample in a large-scale training dataset. However, considering the time and labor costs of labeling, it becomes unfeasible to annotate all collected samples. Consequently, a data selection strategy is needed to identify a subset of unlabeled data for labeling. The core objective of this work is to address the active learning (AL) problem in multi-task learning (MTL), where the goal is to maximize performance across multiple tasks while concurrently minimizing the need for large amounts of labeled training data.

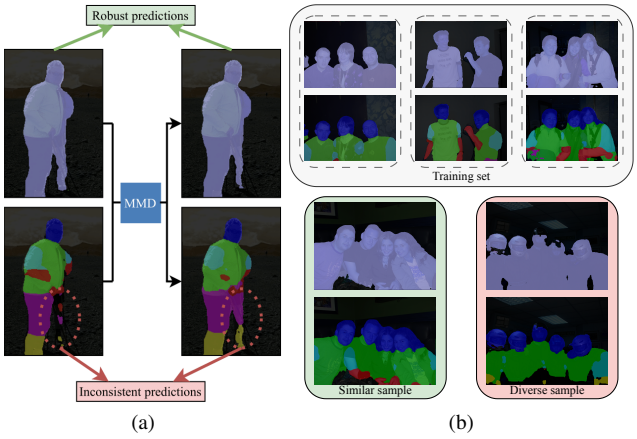


Figure 1. (a) Initial and refined predictions of a two-stage MTL network with an MMD-block [30]. The tasks with inconsistent predictions lack robustness and need to be further trained. (b) Each sample is visualized as its task predictions. *Diverse* sample includes an interesting scenario with people wearing motorcyclist helmets. In contrast, the *similar* sample consists of a group facing the camera, which is similar to samples in the training set.

Inconsistency-based single-task AL strategies [11, 32] provide a solution where the selection is based on the inconsistency in the predictions of a neural network under different augmentations applied to the same input image. Building upon this principle, we propose leveraging the inherent inconsistency within multi-task refinement networks [28, 30]. As depicted in Fig. 1a, initial predictions of these networks are refined using a multi-modal distillation (MMD) block that shares information between the task pre-

dictions. We observe that more challenging samples result in higher levels of inconsistency. Therefore, we formulate our AL selection strategy by quantifying the inconsistency between these refinements and the initial predictions.

Diversity-based AL strategies [24] aim to curate a diverse training set, where the diversity is defined by the distance between feature embeddings extracted from a backbone neural network. However, adapting this technique to the high-dimensional intermediate feature maps, typically found after the backbone layers of multi-task networks, and further extending it to capture diverse scenarios across multiple tasks is a non-trivial challenge yet to be addressed. We introduce a novel solution using intermediary feature maps obtained from each task-specific head to capture task-specific diversity. We further distill the multi-task representation into a compact feature vector using an auto-encoder architecture [13]. This novel approach allows us to encapsulate task-specific information across all tasks within a singular vector, effectively defining a measure of diversity. As depicted in Fig. 1b, this strategy allows us to capture distinctive task-specific scenarios. For instance, the motorcycle helmet introduces a unique challenge for head segmentation, and learning on these task-specific diverse samples would improve the performance of the network on these samples in further iterations. Notably, given the visual similarities of scenes, conventional strategies might fail to capture such task-specific information effectively.

In multi-task active learning (MTAL), once samples are selected, the labeling process occurs simultaneously across all tasks. However, when the selection scores from multiple tasks are aggregated, a sample having high priority for one task might not necessarily be equally beneficial for the learning process of another task. This scenario contradicts the principles of single-task AL, which centers on both the selection of samples for labeling and the ordering of samples based on their informativeness in the learning process. Training with low informative samples could diminish the impact of high informative samples, potentially leading to overfitting on well-learned, robust predictions. To counteract this issue inherent in multi-task AL, we present a novel loss weighting strategy designed to preserve the effect of individual AL selections for each task by scaling the impact of samples separately based on the corresponding selection score. This strategy ensures that the benefits of single-task AL selection are preserved while concurrently selecting and labeling samples for all tasks.

Our main contributions are summarized as follows:

- We propose an inconsistency-based selection strategy that leverages the inconsistencies between initial and refined task predictions in multi-task refinement networks.
- We present a novel feature embedding approach tailored for diversity-based AL that captures the task-specific characteristics and condenses them into a unified feature vector.
- We introduce a loss weighting strategy to adjust the impact of samples based on their selection scores, effectively simulating the dynamics of single-task AL during training of a multi-task network.

2. Related work

2.1. Uncertainty-based active learning

Uncertainty-based AL methods rank the pool of unlabeled samples using an informativeness score to select the most informative samples for labeling. One common approach is constructing a *committee of models*, which makes multiple predictions for the same input sample. The informativeness of a sample is then measured based on the level of disagreement or inconsistency among the predictions from these committee members. In epistemic uncertainty-based methods, a committee is formed by training multiple models with different initial random weights [4] or through dropout layers, producing diverse predictions in each forward pass [20]. Inconsistency-based methods [8, 11, 32] generate diverse predictions by applying different augmentations to the same input sample. Golestaneh *et al*. [11] utilize horizontal flipping and calculate inconsistency using KL-divergence between predictions from the original and flipped images, resulting in a selection score suitable for dense vision tasks like semantic segmentation. Yoo *et al*. [31] proposed a task-agnostic loss-based selection strategy. They integrate an extra head within the network to predict the target loss for unlabeled samples, and the predicted loss serves as the selection criterion.

Extending the single-task AL to the multi-task setting has been primarily explored within the natural language processing domain. Reichart and Rappoport [23] introduced combining ranks of single-task selection scores by aggregating them into a multi-task score used for ranking samples for selection. Ikhwantri et al. [14] selected a random task at each AL iteration and used its single-task selection score to guide the selection of samples for labeling. Recently, Durasov *et al*. [6] leveraged single-task uncertainty scores to pick a subset of tasks for which labels are annotated. Despite these advancements, a comprehensive multi-task selection strategy that considers task interactions and investigates multi-task diversity remains unexplored.

2.2. Diversity-based active learning

In diversity-based AL [1, 7, 24, 26], the goal is to curate a diverse training set through the selection of representative samples that span the entirety of the input space. In the pioneering Core-Set approach, introduced by Sener *et al*. [24], diversity is defined as the distance between intermediate

network features extracted for each image. These features are extracted from the penultimate layer of the image classification network. Building upon this, Hekimoglu *et al*. [12] adapted the core-set diversity to object detection, demonstrating that using task-specific features improves the selection performance. The CDAL strategy, introduced by Agarwal et al. [1], leverages contextual diversity relative to predicted classes and has been applied to object detection and semantic segmentation. Conversely, adversarial diversity-based approaches [7, 26] involve training a discriminator to predict whether a given sample belongs to the labeled training set. Samples with lower similarity to the labeled set are selected for labeling in the next AL iteration.

To our knowledge, prior work has not extended diversity-based AL methods to the multi-task setting. Our work seeks to bridge this gap by investigating how task-specific features can be extracted for multiple tasks and combined into a unified diversity-based selection strategy.

2.3. Multi-task learning

Recent works on MTL can be categorized into two categories: network architectures and loss-weighting strategies.

A common multi-task architecture for computer vision applications involves a global feature extractor, followed by task-specific output heads for each task [5, 17, 18, 35, 36]. In cross-talk networks [21], instead of a single shared feature extractor, "cross-talk" allows for information flow between parallel layers of individual task-specific networks. Furthermore, prediction distillation architectures [28, 30, 34] generate initial task predictions, then, using task interactions, refine these predictions in the final outputs. As introduced by Xu *et al*. [30], PAD-Net operates in two stages. In the initial stage, a backbone followed by task-specific heads produces initial task predictions. Then, the second stage uses an MMD block on these predictions, which distills information from other tasks and enforces task consistency to enhance the final refined predictions.

Strategies for MTL training balance the joint learning of tasks by a weighing mechanism for combining the loss of individual tasks. Kendall *et al*. [15] leveraged the homoscedastic uncertainty to balance the single-task losses. Similarly, in Dynamic Weight Averaging (DWA), introduced by Liu *et al*. [17], the learning progression across tasks is balanced through a weighing term based on the change of loss for each task. These methods weigh the loss of each task by a quantity, which remains constant across diverse input samples for a single task. In contrast, our proposal weights each sample during training based on the associated task uncertainty. With this, we simulate an AL selection for each task individually and bridge the gap between single-task and multi-task AL.

3. Methodology

The aim of this study is to tackle the AL problem in the context of multi-task networks. We adopt a pool-based AL framework, where we iteratively select samples from a large pool of unlabeled data denoted as U. The selected samples are subsequently labeled by an annotator and then integrated into the training dataset T, for further training in the following iterations. This process continues iteratively, with the aim of refining the multi-task network's performance over time. In our notation, we denote a sample pair as (x, y), where x is an image, and y is the collection of multi-task labels $\{y_k \mid k \in K\}$ with ground-truth annotations for each one of the $k \in K$ tasks.

Our proposed framework can be integrated into any MTL architecture. We illustrate our approach using the two-stage PAD-Net architecture [30] as our multi-task network. PAD-Net consists of two prediction stages. The first stage generates predictions for each task using a shared backbone and task-specific heads. In the second stage, these predictions are refined through an MMD block, which uses an attention-guided message-passing mechanism between the initial predictions to fuse relevant information across tasks.

In Algorithm 1, we present the steps for a single AL iteration in our proposed framework. For every unlabeled sample, we assign two selection scores: a task-inconsistency score (Sec. 3.1.1) and a diversity score (Sec. 3.1.2). Then, we combine these scores into a single selection score (Sec. 3.1.3). Next, we rank the samples based on this combined score and select the top B samples, where B represents the AL budget for that iteration. Subsequently, the newly labeled samples are incorporated into the labeled training set, which is used to train the network Φ in the following iteration. During the training phase, we scale the multi-task weight of each sample, as explained in Sec. 3.2, to imitate single-task AL selection.

Algorithm 1 Single iteration of our AL framework

Require: unlabeled set U, training set T, budget B, network Φ, set of tasks K
for $u \in U$ **do** (Sec. 3.1)
  Compute inconsistency score $s_k(u)$ (Eq. (2))
  Compute diversity score $div(u)$ (Eq. (3))
  Compute combined score $s(u)$ (Eq. (4))
end for
$S \leftarrow$ Select top B samples from U based on $s(u)$
Acquire labels Y_S for samples in S
Augment training set: $T \leftarrow T \cup \{S, Y_S\}$
Train Φ using Eq. (5) on T scaled by s_k (Sec. 3.2)

3.1. Selection strategy

3.1.1 Inconsistency-based selection score

Recent two-stage multi-task networks, incorporating multi-modal distillation layers like PAD-Net [30], have demonstrated the potential of leveraging task interactions to refine the initial task predictions. For instance, if the boundary of a semantic segmentation prediction does not align with the boundary identified by an edge detection task, these tasks exchange information via an attention mechanism, leading to improved predictions in the second stage.

We base our inconsistency-based selection score on the robustness of the predictions of a task before and after the refinement iteration. When the initial and refined predictions for a task show significant disparity, indicating a lack of robustness, it suggests that the task is not certain for that sample. Consequently, the respective task head could benefit from learning from that sample. With this motivation, we propose constructing a *committee of refined predictions* and measuring task robustness by quantifying disagreement among these committee predictions.

To construct the *committee of refined predictions*, we extend the PAD-Net architecture by integrating pairwise task interactions. Specifically, the predictions of each task are refined by incorporating predictions from all other tasks through a two-task MMD block. Since the MMD block is based on a spatial attention mechanism, the interactions are learnable, enabling the network to disregard irrelevant task interactions. These additional blocks are only utilized for selection purposes and introduce no additional computational overhead during inference since we only utilize the layers from the original PAD-Net architecture. We refer to the supplementary for more implementation details.

In Fig. 2, we demonstrate the score calculation process using the depth estimation task as an illustrative example. For each task, we refine the initial prediction by incorporating predictions from all other tasks, yielding $K-1$ refined predictions. Together with the initial prediction, this results in a committee of K predictions. To quantify the disagreement among these committee predictions, we utilize the corresponding task-specific loss function. Using the loss allows for an easily applicable and task-agnostic approach.

Formally, for task k, we calculate the inconsistency score as the maximum loss between the initial task prediction $\hat{y}_k(x)$ and the refined task predictions $\hat{y}_{kj}(x)$ for all other tasks $j \in K, j \neq k$:

$$\hat{s_k}(x) = \max_{j \in K, j \neq k} L_k(\hat{y}_k(x), \hat{y}_{kj}(x)) \tag{1}$$

To ensure comparability across tasks with varying score ranges, we normalize the task inconsistency scores within the [0,1] range as follows:

$$s_k(x) = \frac{\hat{s_k}(x) - s_k^{min}}{s_k^{max} - s_k^{min}} \tag{2}$$

where s_k^{min} and s_k^{max} respectively represent the minimum and maximum values of $\hat{s_k}$ for task k across all samples.

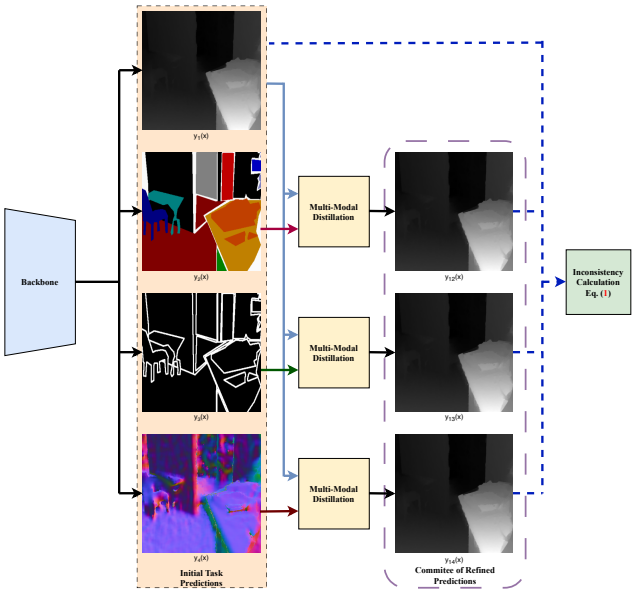


Figure 2. Visualization of the inconsistency score calculation. For each task, depth estimation is shown as an example, initial prediction is refined by pairwise initial predictions of other tasks to construct a *committee of refined predictions*. The inconsistency within this committee defines the inconsistency-based selection score.

3.1.2 Task-specific diversity score

In the Core-Set approach [24], diversity is defined by the distance between images, where the distance is computed as the Euclidean distance between the intermediate feature vectors of the penultimate layer in an image classification network. One straightforward extension of this approach to the multi-task setting would be to employ the distance between the intermediate feature maps obtained after the backbone before the task-specific heads. However, this high-dimensional representation makes the distance metric unreliable. Furthermore, recent findings indicate that generalized backbone features do not effectively diversify task-specific scenarios. Instead, employing task-specific features is shown to yield better overall AL selection performance [12]. Consequently, within the multi-task context, the challenge arises of capturing task-specific features while summarizing them into a more compact representation.

In Fig. 3, we provide an overview of our diversity-based strategy. To capture task-specific feature representations,

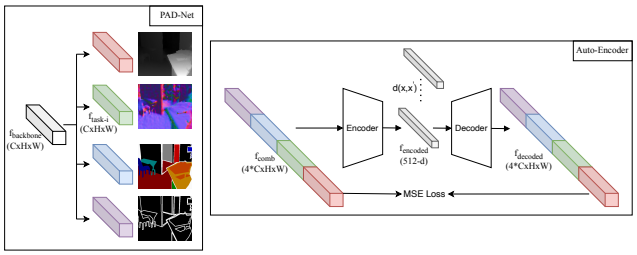


Figure 3. Illustration of the proposed feature extraction strategy. We combine the features from task-specific heads in the PAD-Net architecture and use an auto-encoder to obtain a representative feature vector. Then, the distance between samples d is defined by the distance between the encoded feature embeddings.

we extract intermediary feature maps obtained from the final layers of each task-specific head. We then concatenate these feature maps to form a unified representation that captures the relevant task-specific features across all tasks. However, this combined representation is high-dimensional and may contain some task-related features redundantly. For instance, information related to an edge in both edge detection and semantic segmentation can be distilled into a more summarized representation. Therefore, to address the challenges of dimensionality reduction and summarization of information from different task-specific features, we utilize an auto-encoder architecture.

We encode the concatenated feature maps into an N-dimensional feature vector and decode them back to the same representation. Empirically, we determined that a 512-dimensional embedding offers optimal results for our diversity representation. A detailed ablation study is provided in the supplementary. To train the auto-encoder, we use a mean-square error loss between the concatenated feature map and the decoded output. The auto-encoder is trained separately from the multi-task network after completing the AL training phase for the multi-task network.

During the selection phase, we use the encoded features as our diversity embeddings. These embeddings encapsulate task-specific features and are compactly summarized in a feature vector. The diversity-based selection score for a sample is defined by the distance between its embedding and the closest sample in the training set, as follows:

$$div(x) = \min_{x' \in T} d(x, x') \tag{3}$$

where x represents an unlabeled sample, while x' iterates over all samples in the training set T. For the choice of the distance measure d, we conduct an ablation study in the supplementary with various metrics and empirically find that the Euclidean distance produces the best performance.

3.1.3 Combined selection score

To effectively address task inconsistency and sample diversity during selection, we combine the two selection scores to rank the samples for labeling. For the task inconsistency-based score, we prioritize the sample with the highest inconsistency by selecting the maximum of the individual task scores. This score is then multiplied by the diversity-based score. Formally, the combined score is represented as:

$$s(x) = div(x) \cdot \max_{k \in K} s_k(x) \tag{4}$$

3.2. Training strategy

Within an MTAL framework, samples are assumed to be labeled for all tasks simultaneously. However, the selection scores we obtain in Eq. (1) are specific to individual tasks. Consequently, a sample could be selected for its high inconsistency score for one task despite having a significantly lower score for another. Such a selection is counter-intuitive in the context of AL. Training with samples of lower priority reduces the significance of the high-priority samples and poses a risk of overfitting on well-learned samples.

To keep the principles of single-task AL selection in the multi-task setting, we propose an adaptive training strategy. This approach prioritizes samples as if operating under a single-task AL scenario. To achieve this, we scale the training influence of each sample based on its task-inconsistency score. This strategy ensures that the network does not overly learn from samples with low inconsistency while giving greater importance to those with high inconsistency.

To implement this strategy, we modify the conventional weighted multi-task loss [27] by the normalized task-inconsistency selection score Eq. (2). The modified loss function is then defined as:

$$L(x, y) = \sum_{k \in K} s_k(x) \cdot w_k \cdot L_k(\hat{y}_k, y_k) \tag{5}$$

Here, w_k represents the task-specific weight parameter (such as DWA [17] or homoscedastic uncertainty [15]), which controls the relative influence of each task during training. Meanwhile, we use s_k to serve as a sample-specific scaling factor. We note that while w_k remains constant for different input samples of the same task, s_k dynamically scales the loss for each sample to mimic the single-task AL process.

4. Experiments

4.1. Experimental setup

Datasets. We assess the effectiveness of our approach using two publicly available datasets: PASCAL [10] and

NYUD-v2 [25]. For the PASCAL dataset, we follow the split provided by PASCAL-Context [2], resulting in 4,998 training and 5,105 validation images. These are annotated for tasks including semantic segmentation, human part segmentation, and semantic edge detection. Following the extension from [27], we also consider surface normals prediction and saliency detection tasks. NYUD-v2 dataset [25] contains 795 training and 654 validation images that depict indoor scenes. Labels are available for semantic segmentation and monocular depth estimation. Following [27, 30], we also include surface normals labels estimated from the depth maps.

Evaluation metrics. We use the *mean intersection over union (mIoU)* (↑) [9] as the evaluation metric for semantic segmentation, saliency estimation, and human part segmentation, *the optimal dataset F-measure (odsF)* (↑) [19] for edge detection, *root mean square error (rmse)* (↓) for depth estimation and *mean error (mErr)* (↓) in the predicted angles to evaluate the surface normals. For the multi-task performance we use the *multi-task loss* (↓).

Network details. We adopt the multi-task architecture PAD-Net, introduced by Xu *et al*. [30] with HRNet [29] backbone, and further extend it with our AL framework. For PASCAL, we follow [28] and use stochastic gradient descent with momentum 0.9 with a batch size of 8 and a learning rate of 1e-2. The multi-task weights w_k are set as follows: 1.0 for semantic segmentation, 2.0 for human parts segmentation, 5.0 for saliency, 10.0 for normals estimation, and 50.0 for edge detection. For NYUD-v2, we utilize an Adam optimizer with an initial learning rate of 1e-4 and a batch size of 6. The multi-task weights are 1.0 for semantic segmentation, 1.0 for depth estimation, and 10.0 for normals estimation.

Active learning details. To initiate the active learning process, we treat all the samples in the training split as the unlabeled pool and create an initial labeled training set by randomly selecting 10% of available samples. We start all the baselines from the same network trained with this initial dataset. At each AL iteration, we add 10% more samples from the remaining unlabeled pool to the training set, guided by the corresponding selection strategy. To simulate the labeling process, we use the already available annotations. We employ a continuous training strategy, where in each AL iteration, the network is initialized with the best-performing weights from the previous iteration. During each training iteration, the network is trained for 20 epochs.

We report the mean of the respective metric on the validation sets for each experiment over three independent runs with different random selections of the initial labeled pool. We also report the numerical values and variances across these three runs in the supplementary. All experiments are conducted using two Tesla V100 GPUs.

4.2. Comparison with state-of-the-art

Baselines. We compare the effectiveness of our MTAL method against several baseline methods.

- *Random.* This baseline mimics passive learning, where each sample receives a selection score generated from a uniform distribution.
- *Core-Set [24].* We use the feature map obtained after the backbone network and before the task-specific head as the diversity embeddings. We use the k-Center-Greedy solver from [24], where we define the distance to be the Euclidean distance between the feature maps of different samples.
- *LL4AL [31].* We incorporate the loss prediction strategy proposed by Yoo *et al*. [31]. We extend the network architecture with a loss prediction module for each task, enabling it to predict the loss of each sample. Samples with the highest predicted total loss are selected for labeling.
- *EquAL [11].* To compare against single-task selection, we employ the SOTA AL for segmentation strategy EquAL [11]. Since there are no specific AL strategies for the other tasks, we adapt this baseline separately for each task. We select by only considering the score of a single task. For example, we denote the corresponding selection strategy as *EquAL-DE* for depth estimation.
- *RBAL [23].* This baseline ranks each task based on single-task uncertainty (for our experiments through EquAL). These ranks are then summed per sample to generate a combined ranking score.
- *PartAL [6].* This method assigns an uncertainty score per task and, for each sample, only partially labels the relevant tasks instead of labeling all tasks.

We also include the AP of a "fully-trained" network, which is trained on the entire training set, to provide a reference point for the potential of our network. For more implementation details, we direct readers to the supplementary.

Results on PASCAL. In Fig. 4a, we compare our method against the baselines on the PASCAL dataset, using the multi-task loss. After the initial AL cycle, corresponding to 20% of labeled samples, our method achieves a 3.9% lower loss than other baselines. As the number of actively selected labels increases, for example, using 60% of all available data, with 50% actively labeled, our method demonstrates a 21.8% reduction in loss relative to the *Random* baseline and an 8.5% reduction in loss in contrast to the second-best baseline, *PartAL*. Notably, due to the continuous training strategy, several baselines exceed the "fully-trained" performance as they start training with

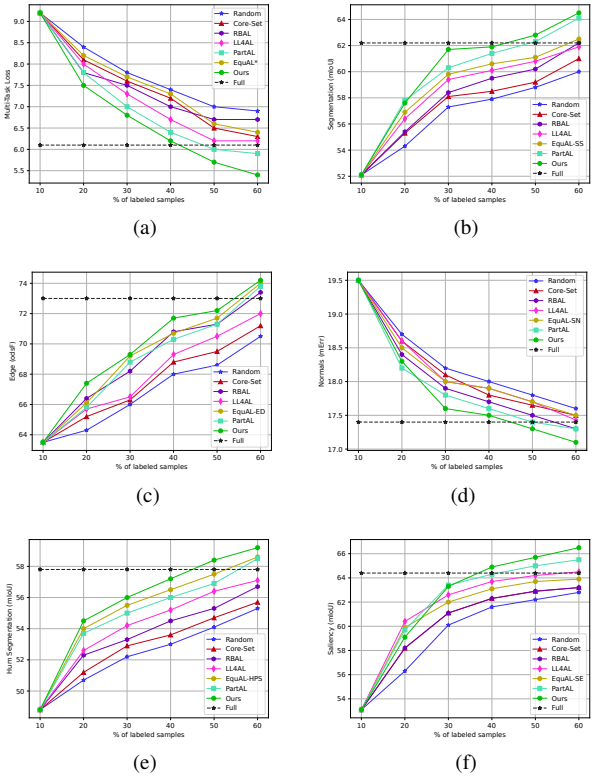


Figure 4. Comparison of our proposed method with SOTA AL methods on the PASCAL dataset. Lines indicate the average results over three trials. EquAL* represents the single-task EquAL baseline with the lowest final MTL loss across all tasks.

high-priority samples and iteratively refine their networks with newly added samples. Our method manages to reach the loss of a fully-trained network while using only 42% of the data. In contrast, *PartAL* requires 48% of the data to achieve similar performance, representing a 6% saving in data utilization.

Furthermore, in Figures 4b, 4c, 4d, 4e, and 4f, we present the individual AL graphs for each task. Our method consistently outperforms both multi-task and single-task AL strategies across all tasks. This highlights the efficacy of our approach in terms of maintaining single-task AL performance while leveraging the insights derived from the proposed multi-task inconsistency and diversity selection scores.

Results on NYU. We present the multi-task loss for the NYUv2 dataset for various baselines in Fig. 5. Similar to PASCAL results, our method consistently outperforms the other multi-task and single-task baselines across all tasks. Our method reaches the loss of the fully-trained network using 50% of available data, corresponding to a 10% data savings rate, compared to the second-best method, PartAL.

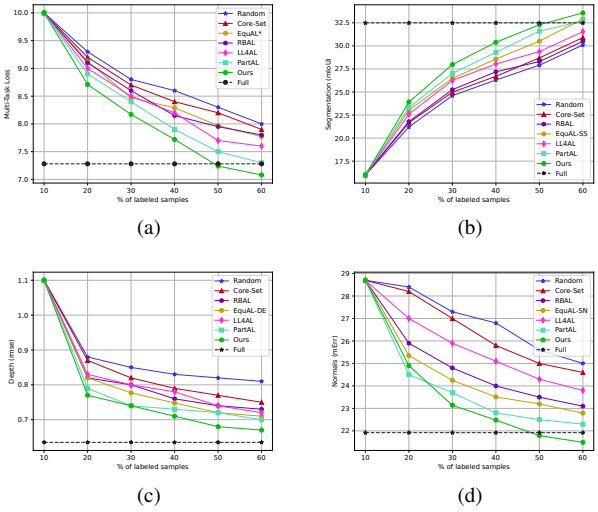


Figure 5. Comparison of our proposed method with SOTA AL methods on the NYU dataset. Lines indicate the average results over three trials. EquAL* represents the single-task EquAL baseline with the lowest final MTL loss across all tasks.

4.3. Qualitative comparison

Fig. 6 qualitatively compares network predictions trained with 60% of the data selected by our proposed method and the PartAL baseline. Our method results in better segmentation of the intricate details, particularly in some challenging scenarios, such as a bride's wedding dress and a sailor's hat. This success can be attributed to the effectiveness of our selection strategy, which can identify such diverse scenarios. Also, due to our inconsistency-based selection strategy, we showcase better inconsistency between the two tasks, apparent in segmenting the corner of the sailor's hat.

4.4. Ablation study

Effect of committee size. One of the design choices in our approach is the utilization of pairwise task interactions to construct a committee with K predictions. To investigate the influence of committee size on selection performance, we experiment with variations involving triplet and quadruple connections. Specifically, for the PASCAL dataset with five tasks, the initial prediction constitutes one committee member per task. Adding pairwise connections expands the committee to five predictions (*C5*). Incorporating triplet connections introduces six additional members (*C11*), and quadruple connections contribute another four members (*C15*). Results from this ablation study, presented in Fig. 7a, indicate that the performance gain after five members (*C5*) is minimal. Therefore, to reduce complexity, we use only pairwise connections.

Different variants of diversity. We compare our approach to various alternatives diversity-based strategies

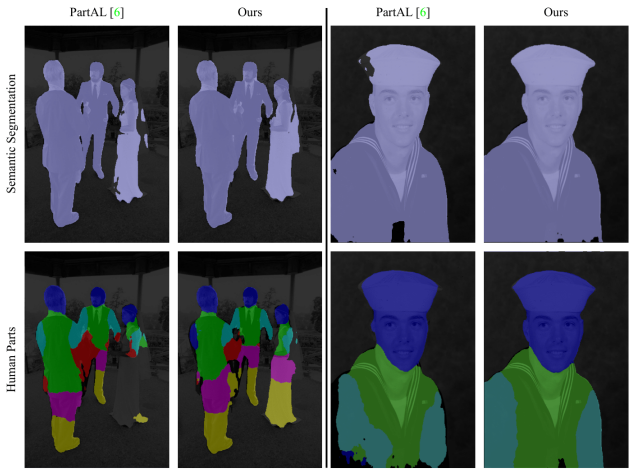
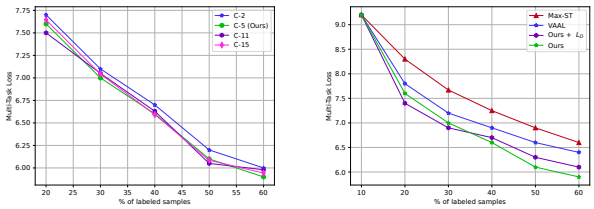


Figure 6. Qualitative comparison of the network predictions trained with 60% of the data selected by our proposed method and the PartAL [6] baseline, on the PASCAL dataset.



(a) (b)

Figure 7. Ablation study on: (a) effect of committee size on the inconsistency-based selection performance, (b) different variants of the diversity-based selection strategy.

in Fig. 7b. We compare against using the maximum of single-task diversity scores *Max-ST*, where for each task, we use the feature maps in the penultimate layers of the corresponding task-specific head. Then, we take the maximum task-specific distances to represent the diversity score. Additionally, we compare against an adversarial diversity-based approach *VAAL*, with the same concatenated feature maps we feed to our autoencoder as inputs to the VAAL structure proposed by Sinha *et al*. [26]. We also introduce a variant of our approach $Ours + L_D$, incorporating the discriminator loss from VAAL into our auto-encoder learning scheme. The results indicate that using combined multi-task features for diversity (*VAAL*, *Ours*) reaches higher performance. Moreover, adding the discriminator loss increases the performance in early cycles but is surpassed after some iterations. This indicates that the discriminator loss can be beneficial in early cycles in identifying outliers. However, as the distribution gap between the labeled and the unlabeled set closes, discriminating between two close distributions becomes more ambiguous, and the performance gain of L_D is degraded.

Comparison of loss weighting strategies. In Tab. 1, we evaluate two performance of two sample-specific loss weighting strategies in conjunction with two task-specific loss weighting strategies. We compare the *Fixed* task-specific weights, using the weights detailed in Sec. 4.1 against DWA [17]. Additionally, we compare our sample-specific weighting method, which uses the inconsistency, against weighting by the loss of a sample, which we calculate using ground-truth annotations after labeling. Results demonstrate that the best MTL performance is achieved when combining the *Fixed* weighting strategy with our proposed inconsistency-based scaling method. Using loss as the scaling factor results in lower performance regardless of the task-specific weighting strategy. This suggests that inconsistency is a more reliable proxy for weighting the samples during the multi-task training phase.

Weights		Seg.	Depth	MTL
Task w_k	Sample $s_k(x)$	IoU	rmse	Loss
DWA [17]	-	32.1	0.74	7.68
	Loss	32.5	0.72	7.37
	Incons.	32.9	**0.65**	7.25
Fixed	-	32.7	0.78	7.63
	Loss	32.3	0.77	7.55
	Incons.	**33.6**	0.67	**7.08**

Table 1. Ablation study of different MTL loss weighting strategies on the *NYUD-v2*. w_k and $s_k(x)$ represent the task and sample-specific loss weights in Eq. (5). "-" denotes no scaling, *Loss* denotes using the ground-truth loss for scaling and *Incons.* denotes using our proposed sample inconsistensy score.

5. Conclusion

We introduced a novel AL strategy for multi-task networks for vision tasks. Our selection strategy consists of a task-inconsistency-based selection score and a multi-task-specific diversity score. We further proposed a novel multi-task learning training strategy to simulate the effects of single-task AL in the multi-task setting. Our experiments on two multi-task datasets demonstrate the effectiveness of our approach, as it reduces the fully-trained loss by 2.8% and achieves 10% fewer annotations than the state-of-the-art baseline. One interesting future direction is to investigate integrating semi-supervised learning techniques, such as consistency regularization, into our framework to further reduce the annotation requirements.

References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *ECCV*, 2020. 2, 3

[2] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014. 6

[3] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, 2020. 1

[4] Kashyap Chitta, José M Álvarez, Elmar Haussmann, and Clément Farabet. Training data subset search with ensemble active learning. *T-ITS*, 2021. 2

[5] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3150–3158, 2016. 3

[6] Nikita Durasov, Nik Dorndorf, and Pascal Fua. Partal: Efficient partial active learning in multi-task visual settings. *arXiv preprint arXiv:2211.11546*, 2022. 2, 6, 8

[7] Sayna Ebrahimi, William Gan, Dian Chen, Giscard Biamby, Kamyar Salahi, Michael Laielli, Shizhan Zhu, and Trevor Darrell. Minimax active learning. *arXiv preprint arXiv:2012.10467*, 2020. 2, 3

[8] Ismail Elezi, Zhiding Yu, Anima Anandkumar, Laura Leal-Taixe, and Jose M Alvarez. Not all labels are equal: Rationalizing the labeling costs for training object detection. In *CVPR*, 2022. 2

[9] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 6

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 5

[11] S Alireza Golestaneh and Kris M Kitani. Importance of self-consistency in active learning for semantic segmentation. In *BMVC*, 2020. 1, 2, 6

[12] Aral Hekimoglu, Michael Schmidt, Alvaro Marcos-Ramiro, and Gerhard Rigoll. Efficient active learning strategies for monocular 3d object detection. In *IV*, 2022. 3, 4

[13] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 2006. 2

[14] Fariz Ikhwantri, Samuel Louvan, Kemal Kurniawan, Bagas Abisena, Valdi Rachman, Alfan Farizki Wicaksono, and Rahmad Mahendra. Multi-task active learning for neural semantic role labeling on low resource conversational corpus. In *ACLW*, 2018. 2

[15] Alex Kendall, Yarin Gal, and Roberto 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*, pages 7482–7491, 2018. 3, 5

[16] Ryohei Kuga, Asako Kanezaki, Masaki Samejima, Yusuke Sugano, and Yasuyuki Matsushita. Multi-task learning using multi-modal encoder-decoder networks with shared skip connections. In *ICCVW*, 2017. 1

[17] 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. 3, 5, 8

[18] 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 *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939, 2018. 3

[19] David R Martin, Charless C Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *TPAMI*, 2004. 6

[20] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *ICRA*, 2018. 2

[21] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, 2016. 3

[22] Arsalan Mousavian, Hamed Pirsiavash, and Jana Košecká. Joint semantic segmentation and depth estimation with deep convolutional networks. In *3DV*, 2016. 1

[23] Roi Reichart, Katrin Tomanek, Udo Hahn, and Ari Rappoport. Multi-task active learning for linguistic annotations. In *ACL*, 2008. 2, 6

[24] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. 2, 4, 6

[25] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*. Springer, 2012. 6

[26] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *ICCV*, 2019. 2, 3, 8

[27] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *TPAMI*, 2021. 5, 6

[28] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *ECCV*, 2020. 1, 3, 6

[29] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2020. 6

[30] 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 *CVPR*, 2018. 1, 3, 4, 6

[31] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *CVPR*, 2019. 2, 6

[32] Weiping Yu, Sijie Zhu, Taojiannan Yang, and Chen Chen. Consistency-based active learning for object detection. In *CVPR*, 2022. 1, 2

[33] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 1

[34] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, 2019. 3

[35] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*. Springer, 2014. 3

[36] Xiangyun Zhao, Haoxiang Li, Xiaohui Shen, Xiaodan Liang, and Ying Wu. A modulation module for multi-task learning with applications in image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–416, 2018. 3