MTI-Net: Multi-Scale Task Interaction Networks for Multi-Task Learning

Simon Vandenhende[1] Stamatios Georgoulis[2] Luc Van Gool[1,2]

[1]KU Leuven/ESAT-PSI [2]ETH Zurich/CVL

Abstract. In this paper, we argue about the importance of considering task interactions at multiple scales when distilling task information in a multi-task learning setup. In contrast to common belief, we show that tasks with high affinity at a certain scale are not guaranteed to retain this behaviour at other scales, and vice versa. We propose a novel architecture, namely MTI-Net, that builds upon this finding in three ways. First, it explicitly models task interactions at every scale via a multi-scale multi-modal distillation unit. Second, it propagates distilled task information from lower to higher scales via a feature propagation module. Third, it aggregates the refined task features from all scales via a feature aggregation unit to produce the final per-task predictions.
Extensive experiments on two multi-task dense labeling datasets show that, unlike prior work, our multi-task model delivers on the full potential of multi-task learning, that is, smaller memory footprint, reduced number of calculations, and better performance w.r.t. single-task learning.

Keywords: Multi-Task Learning, Multi-Scale Processing, Feature Distillation

1 Introduction and prior work

The world around us is flooded with complex problems that require solving a multitude of tasks concurrently. An autonomous car should be able to detect all objects in the scene, localize them, understand what they are, estimate their distance and trajectory, etc., in order to safely navigate itself in its surroundings. In a similar vein, an intelligent advertisement system should be able to detect the presence of people in its viewpoint, understand their gender and age group, analyze their appearance, track where they are looking at, etc., in order to provide personalized content. The examples are countless. Understandably, this calls for efficient computational models in which multiple learning tasks can be solved simultaneously.

Multi-task learning (MTL) [2,38] tackles this problem. Compared to the single-task case, where each individual task is solved separately by its own network, multi-task networks theoretically bring several advantages to the table. First, due to their layer sharing, the resulting memory footprint is substantially reduced. Second, as they explicitly avoid to repeatedly calculate the features in the shared layers, once for every task, they show increased inference speeds. Most

importantly, they have the potential for improved performance if the associated tasks share complementary information, or act as a regularizer for one another. Evidence for the former has been provided in the literature for certain pairs of tasks, e.g. detection and classification [10,35], detection and segmentation [7,13], segmentation and depth estimation [8,48], while for the latter recent efforts point to that direction [43].

Motivated by these observations, researchers started designing architectures capable of learning shared representations from multi-task supervisory signals. Misra et al. [32] proposed to use "cross-stitch" units to combine features from multiple networks to learn a better combination of shared and task-specific representations. Kokkinos [20] introduced a multi-head architecture called UberNet that jointly handles as many as seven tasks in a unified framework, which can be trained end-to-end. Doersch and Zisserman [6] exploited multiple self-supervised tasks in order to train a single visual representation via a "lasso" regularization scheme. Zamir et al. [50] proposed to model the structure of the visual tasks' space by finding transfer learning dependencies across a dictionary of twenty six tasks. Despite the progress reported by these or similar works [40,28,33,26,45], the joint learning of multiple tasks can lead to single-task performance degradation if information sharing happens between unrelated tasks. The latter is known as *negative transfer* [53], and has been well documented in [20], where an improvement in estimating normals leads to a decline in object detection, or in [13] where the multi-task version underperforms the single-task ones.

To remedy this situation, a group of methods carefully balance the losses of the individual tasks, in an attempt to find an equilibrium where no task declines significantly. For example, Kendall et al. [17] used the homoscedastic uncertainty of each individual task to re-weigh the losses. Gradient normalization [5] was proposed to balance the losses by adaptively normalizing the magnitude of each task's gradients. Similarly, Sinha et al. [42] tried to balance the losses by adapting the gradients magnitude, but differently, they employed adversarial training to this end. Dynamic task prioritization [11] proposed to dynamically sort the order of task learning, and prioritized 'difficult' tasks over 'easy' ones. Zhao et al. [53] introduced a modulation module to encourage feature sharing among 'relevant' tasks and disentangle the learning of 'irrelevant' tasks. Sener and Koltun [39] proposed to cast multi-task learning into a multi-objective optimization scheme, where the weighting of the different losses is adaptively changed such that a Pareto optimal solution is achieved.

In a different vein, Maninis et al. [30] followed a 'single-tasking' route. That is, in a multi-tasking framework they performed separate forward passes, one for each task, that activate shared responses among all tasks, plus some residual responses that are task-specific. Furthermore, to suppress the negative transfer issue they applied adversarial training on the gradients level that enforces them to be statistically indistinguishable across tasks during the update step.

Note that all aforementioned works so far follow a common pattern: they *directly* predict all task outputs from the same input in one processing cycle (i.e. all predictions are generated once, in parallel or sequentially, and are not refined

afterwards). By doing so, they fail to capture commonalities and differences among tasks, that are likely fruitful for one another (e.g. depth discontinuities are usually aligned with semantic edges). Arguably, this might be the reason for the only moderate performance improvements achieved by this group of works (see [30]). To alleviate this issue, a few recent works first employed a multi-task network to make initial task predictions, and then leveraged features from these initial predictions in order to further improve each task output – in a one-off or recursive manner. In particular, Xu et al. [48] proposed to distil information from the initial predictions of other tasks, by means of spatial attention, before adding it as a residual to the task of interest. Zhang et al. [51] opted for sequentially predicting each task, with the intention to utilize information from the past predictions of one task to refine the features of another task at each iteration. In [52], they extended upon this idea. They used a recursive procedure to propagate similar cross-task and task-specific patterns found in the initial task predictions. To do so, they operated on the affinity matrices of the initial predictions, and not on the features themselves, as was the case before [48,51].

Although better performance improvements have been reported in these works, albeit for specific datasets (see [48]), they are all based on the principle that the interactions between tasks, which are essential in the distillation or propagation procedures described above, only happen at a fixed, local or global, scale[1]. For all we know, however, this is not always the case. In fact, two tasks with high pattern affinity at a certain scale are not guaranteed to retain this behaviour at other scales, and vice versa. Take for example the tasks of semantic segmentation and depth estimation, and consider the case where two cars at different distances are in front of our camera's viewpoint, with one partially occluding the other. Looking at the local scale (i.e. patch level), the discontinuity in depth labels in the region in-between cars suggests that a similar pattern should be present in the semantic labels, i.e. there should be a change of semantic labels in the exact same region, despite the fact that this is incorrect. However, looking at the global scale this ambiguity can be resolved. An analogous observation can be made if we swapped the order of tasks, and went from global to local scale. We conclude that pattern affinities should not be considered at the task level only, as existing works do [48,51,52], but be conditioned on the scale level too (for a more detailed discussion visit Section 2.2).

In this paper, we go beyond these limitations and explicitly consider interactions at separate scales when propagating features across tasks. We propose a novel architecture, namely MTI-Net, that builds upon this idea. Starting from a multi-scale feature representation of the input image, generated from an off-the-shelf backbone network (e.g. [24,46]), we make an initial prediction for each task at each considered scale (four scales in our case). Next, for each task we distill information from other tasks by means of spatial attention to refine the features of the initial predictions. Note that this process happens at each scale

[1] With the exception of [51], where a first attempt for multi-scale processing happens at the decoding stage, in a strict sequential manner. Note that, their approach is only suitable for a pair of tasks, and can not be extended to multi-task learning.

separately in order to capture the unique task interactions that happen at each individual scale, as discussed above. To tackle the limited field-of-view at higher scales of the backbone network, which can hinder task predictions at these scales, we propose to propagate distilled task information from the lower scales. At the final stage, the distilled features of each task from all scales are aggregated to arrive at the final task predictions.

Our contributions are threefold: (1) we propose to explicitly consider multi-scale interactions when distilling information across tasks in multi-task networks; (2) we introduce an architecture that builds upon this idea with dedicated modules, i.e. multi-scale multi-modal distillation (Section 2.1), feature propagation across scales (Section 2.4), and feature aggregation (Section 2.5); (3) we overcome a common obstacle of performance degradation in multi-task networks, and observe that tasks can mutually benefit from each other, resulting in significant improvements w.r.t their single-task counterparts.

2 Method

2.1 Multi-task learning by multi-modal distillation

Visual tasks can be related. For example, they can share complementary information (surface normals and depth can directly be derived from each other), act as a regularizer for one another (using RGB-D images to predict scene semantics [12] improves the quality of the prediction due to the available depth information), and so on. Motivated by this observation, recent MTL methods [48,51,52] tried to explicitly distill information from other tasks, as a complementary signal to improve task performance. Typically, this is achieved by combining an existing backbone network, that makes initial task predictions, with a multi-step decoding process (see Figure 1 (left)).

In more detail, the shared features of the backbone network are processed by a set of task-specific heads, that produce an initial prediction for every task. We further refer to the backbone and the task-specific heads as the *front-end* of the network. The task-specific heads produce a per-task feature representation of the scene that is more task-aware than the shared features of the backbone network. The information from these task-specific feature representations is then combined via a multi-modal distillation unit, before making the final task predictions. As shown in Figure 1, it is possible that some tasks are only predicted in the front-end of the network. The latter are known as auxiliary tasks, since they serve as proxies in order to improve the performance on the final tasks.

Prior works only differ in the way that the task-specific feature representations are combined. PAD-Net [48] distills information from other tasks by applying spatial attention to these features, before adding them as a residual. PAP-Net [52] recursively combines the pixel affinities from these features during the decoding step. Zhang et al. [51] sequentially predict one task in order to refine its features based on the features of the other task.

For brevity, we adopt the following notations. *Backbone features*: the shared features (at the last layer) of the backbone network; *Task features*: the task-

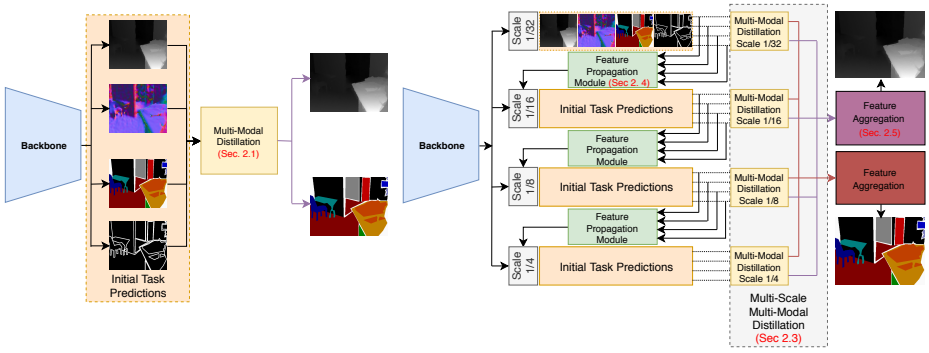


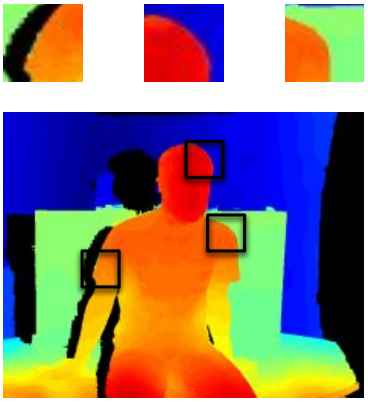
Fig. 1: An overview of different MTL architectures as described in Section 2. (Left) The architecture used in PAD-Net [48] and PAP-Net [52]. Features extracted from a backbone network are used to make initial task predictions. The task features are combined through a distillation unit before making the final task predictions. (Right) The architecture of the proposed MTI-Net. Starting from a backbone that extracts multi-scale features, initial task predictions are made at each scale. The task features are distilled separately at every scale, allowing our model to capture task interactions at multiple scales, i.e. receptive fields. After distillation, the distilled task features from all scales are aggregated to make the final task predictions. To boost performance, we extend our model with a feature propagation mechanism that passes distilled information from lower resolution task features to higher ones.

specific feature representations (at the last layer) of each task-specific head; *Distilled task features*: the task features after multi-modal distillation; *Initial task predictions*: the per-task predictions at the front-end of the network; *Final task predictions*: the network outputs. Note that, backbone features, task features and distilled task features can be defined at a single scale or multiple scales.

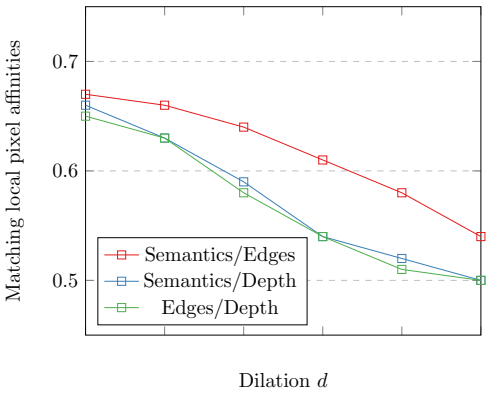
2.2 Task interactions at different scales

The approaches described in Section 2.1 follow a common pattern: they perform multi-modal distillation at a fixed scale, i.e. the backbone features. This rests on the assumption that all relevant task interactions can solely be modeled through a single filter operation with specific receptive field. For all we know, this is not always the case. In fact, tasks can influence each other differently for different receptive field sizes. Consider, for example, Figure 2a. The local patches in the depth map provide little information about the semantics of the scene. However, when we enlarge the receptive field, the depth map reveals a person's shape, hinting at the scene's semantics. Note that the local patches can still provide valuable information, e.g. to improve the local alignment of edges between tasks.

To quantify the degree to which tasks share common local structures w.r.t. the size of the receptive field, we conduct the following experiment. We measure



(a) Three local patches from a depth map. Depending on the patch size, i.e. receptive field, depth information can be utilized differently by other tasks, e.g. semantic segmentation and edges.



(b) To quantify task interactions w.r.t. scale, pixel affinities on the label space of each task, as defined in [52], are calculated in local patches. The correspondences in the affinity patterns between tasks are plotted as a function of the patch size, i.e. kernel dilation.

Fig. 2: Unlike the common belief, in this paper we question whether task interactions remain constant across all scales (see Section 2.2).

the pixel affinity in local patches on the label space of each task, using kernels of fixed size. The size of the receptive field can be selected by choosing the dilation for the kernel. We consider the tasks of semantic segmentation, depth estimation and edge detection on the NYUD-v2 dataset. A pair of semantic pixels is considered similar when both pixels belong to the same category. For the depth estimation task, we threshold the relative difference between pairs of pixels; pixels below the threshold are similar. Once the pixel affinities are calculated for every task, we measure how well similar and dissimilar pairs are matched across tasks. We repeat this experiment using different dilations for the kernel, effectively changing the receptive field. Figure 2b illustrates the result.

A first observation is that affinity patterns are matched well across tasks, with up to 65% of pair correspondence in some cases. This indicates that different tasks can share common structures in parts of the image. This is in agreement with a similar observation made earlier by [52]. A second observation is that the degree to which the affinity patterns are matched across tasks is dependent on the receptive field, which in turn, corresponds to the used dilation. This validates our initial assumption that the statistics of task interactions do not always remain constant, but rather depend on the scale, i.e. receptive field.

Based on these findings, in the next section we introduce a model that distills information from different tasks at multiple scales. By doing so, we are able to capture the unique task interactions at each individual scale, overcoming the limitations of the models described in Section 2.1.

2.3 Multi-scale multi-modal distillation

We propose a multi-task architecture that explicitly takes into account task interactions at multiple scales. Our model is shown in Figure 1 (right). First, an off-the-shelf backbone network extracts a multi-scale feature representation from the input image. Such multi-scale feature extractors have been used in semantic segmentation [36,46,18], object detection [24,46], pose estimation [34,44], etc. In Section 3 we verify our approach using two such backbones, i.e. HRNet [46] and FPN [24], but any multi-scale feature extractor can be used instead.

From the multi-scale feature representation (i.e. backbone features) we make initial task predictions at each scale. These initial task predictions at a particular scale are found by applying a set of task-specific heads to the backbone features extracted at that scale. The result is a per-task representation of the scene (i.e. task features) at a multitude of scales. Not only does this add deep supervision to our network, but the task features can now be distilled at each scale separately. This allows us to have multiple task interactions, each modeled for a specific receptive field size, as proposed in Section 2.2.

Next, we refine the task features by distilling information from the other tasks using a spatial attention mechanism [48]. Yet, our multi-modal distillation process is repeated at each scale, i.e. we apply multi-scale, multi-modal distillation. The distilled task features $F^o_{k,s}$ for task k at scale s are found as:

$$F^o_{k,s} = F^i_{k,s} + \sum_{l \neq k} \sigma\left(W_{k,l,s} F^i_{l,s}\right) \odot \left(W^{'}_{k,l,s} F^i_{l,s}\right), \tag{1}$$

where $\sigma\left(W_{k,l,s} F^i_{l,s}\right)$ returns a per-scale spatial attention mask, that is applied to the task features $F^i_{l,s}$ from task l at scale s. Note that our approach is not necessarily limited to the use of spatial attention, but any type of feature distillation (e.g. squeeze-and-excitation [15]) can easily be plugged in. Through repetition, we calculate the distilled task features at every scale. As the bulk of filter operations is performed on low resolution feature maps, the computational overhead of our model is limited. We make a detailed resource analysis in Section 3.

2.4 Feature propagation across scales

In Section 2.3 actions at each scale were performed in isolation. To sum up, we made initial task predictions at each scale, from which we refined the task features through multi-modal distillation at each individual scale separately. However, as the higher resolution scales have a limited receptive field, the front-end of the network could have a hard time to make good initial task predictions at these scales, which in turn, would lead to low quality task features there. To remedy this situation we introduce a feature propagation mechanism, where the backbone features of a higher resolution scale are concatenated with the task features from the preceding lower resolution scale, before feeding them to the task-specific heads of the higher resolution scale to get the task features there.

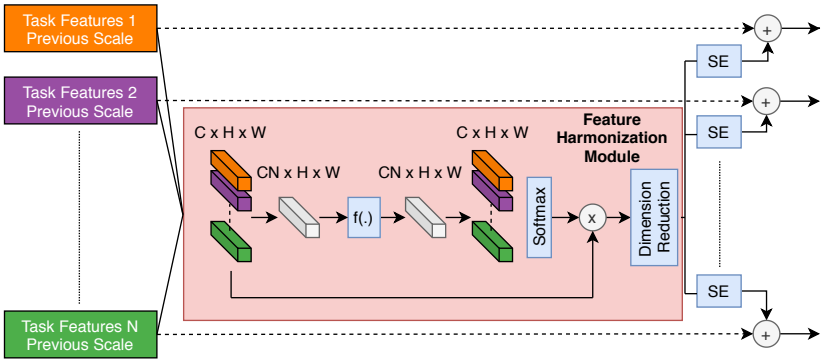


Fig. 3: Our Feature Propagation Module. First, task features from a lower scale are concatenated and mapped to a shared representation by the feature harmonization module. The task features are then refined by extracting information from the shared representation through squeeze-and-excitation (SE) [15], and are added as a residual to the original ones. Finally, these refined task features will be concatenated with the backbone features of the preceding higher scale.

A trivial implementation for our Feature Propagation Module (FPM) would be to just upsample the task features from the previous scale and pass them to the next scale. We opt for a different approach however, and design the FPM to behave similarly to the multi-modal distillation unit of Section 2.3, in order to model task interactions at this stage too. Figure 3 gives an overview of our FPM. We first use a *feature harmonization* block to combine the task features from the previous scale to a shared representation. We then use this shared representation to refine the task features from the previous scale, before passing them to the next scale. The refinement happens by selecting relevant information from the shared representation through a *squeeze-and-excitation* block [15]. Note that, since we refine the features from a single shared representation, instead of processing each task independently as done in the multi-modal distillation unit of Section 2.3, the computational cost is significantly smaller.

Feature harmonization. The FPM receives as input the task features from N tasks of shape $C \times H \times W$. Our feature harmonization module combines the received task features into a shared representation. In particular, the set of N task features is first concatenated and processed by a learnable non-linear function f. The output is split into N chunks along the channel dimension, that match the original number of channels C. We then apply a softmax function along the task dimension to generate a task attention mask. The attended features are concatenated and further processed to reduce the number of channels from $N \cdot C$ to C. The output is a shared representation based on information from all tasks.

Refinement through Squeeze-And-Excitation. The use of a shared representation can degrade performance when tasks are unrelated. We resolve this situation by applying a per-task channel gating function to the shared representation. This effectively allows each task to select the relevant features from

Table 1: Our multi-task learning benchmarks. We predict five tasks on PASCAL. On NYUD-v2 we only consider semantic segmentation and depth, but include edges and normals as auxiliary tasks. Distilled labels are marked with *.

Dataset	Edge	Seg	Parts	Normals	Saliency	Depth
PASCAL	✓	✓	✓	✓*	✓*	
NYUD-v2	✓	✓		✓		✓

the shared representation. The channel gating mechanism is implemented here as a squeeze-and-excitation (SE) block [15]. This is due to the fact that SE has shown great potential in MTL (e.g. [30]), yet other gating mechanisms could be used instead. After applying the SE module, the refined task features are added as a residual to the original task features.

2.5 Feature aggregation

The multi-scale, multi-modal distillation described in Section 2.3 results in distilled task features at every scale. The latter are upsampled to the highest scale and concatenated, resulting in a final feature representation for every task. The final task predictions are found by decoding these final feature representations by task-specific heads again. All implementation details are discussed in Section 3. It is worth mentioning that our model has the possibility to add auxiliary tasks in the front-end of the network, similar to PAD-Net [48]. In our case however, the auxiliary tasks are predicted at multiple scales.

3 Experiments

3.1 Experimental setup

Datasets. We perform our experimental evaluation on the PASCAL [9] and NYUD-v2 [41] datasets. Table 1 contains the tasks that we considered for each dataset. We use the original 795 train and 654 test images for the NYUD-v2 dataset. For PASCAL, we use the split from PASCAL-Context [4] which has annotations for semantic segmentation, human part segmentation and edge detection. We obtain the surface normals and saliency labels from [30], that distilled them from pre-trained state-of-the-art models [1,3].

Implementation details. We build our approach on top of two different backbone networks, i.e. FPN [24] and HRNet [44]. We use the different output scales of the selected backbone networks to perform multi-scale operations. This translates to four scales (1/4, 1/8, 1/16, 1/32). The task-specific heads are implemented as two basic residual blocks [14]. All our experiments are conducted with pre-trained ImageNet weights. We use the L1 loss for depth estimation and the cross-entropy loss for semantic segmentation on NYUD-v2. As in prior

work [19,29,30], the edge detection task is trained with a positively weighted $w_{pos} = 0.95$ binary cross-entropy loss. We do not adopt a particular loss weighing strategy on NYUD-v2, but simply sum the losses together. On PASCAL, we reuse the training setup from [30] to facilitate a fair comparison. We reuse the loss weights from there. The initial task predictions in the front-end of the network use the same loss weighing as the final task predictions. In contrast to [48,52,51], we do not use a two-step training procedure where the front-end is pre-trained separately. Instead, we simply train the complete architecture end-to-end. We refer to the supplementary material for further implementation details. The code will be made available upon publication.

Evaluation metrics. We evaluate the performance of the backbone networks on the single tasks first. The optimal dataset F-measure (*odsF*) [31] is used to evaluate the edge detection task. The semantic segmentation, saliency estimation and human part segmentation tasks are evaluated using mean intersection over union ($mIoU$). We use the mean error ($mErr$) in the predicted angles to evaluate the surface normals. The depth estimation task is evaluated using the root mean square error ($rmse$). We measure the *multi-task learning performance* Δ_m as in [30], i.e. the multi-task performance of model m is defined as the average per-task drop in performance w.r.t. the single-task baseline b:

$$\Delta_m = \frac{1}{T} \sum_{i=1}^{T} (-1)^{l_i} \left(M_{m,i} - M_{b,i}\right) / M_{b,i}, \tag{2}$$

where $l_i = 1$ if a lower value means better for performance measure M_i of task i, and 0 otherwise. The single-task performance is measured for a fully-converged model that uses the same backbone network only for that task.

Baselines. On NYUD-v2, we compare MTI-Net against the state-of-the-art PAD-Net [48]. PAD-Net was originally designed for a single scale, but it is easy to plug-in a multi-scale backbone network and directly compare the two approaches. In contrast, a comparison with [51] is not possible, as this work was strictly designed for a pair of tasks, without any straightforward extension to the MTL setting. Finally, PAP-Net [52] adopts an architecture that is similar to PAD-Net, but the multi-modal distillation is performed recursively on the feature affinities. We chose to draw the comparison with the more generic PAD-Net, since it performs on par with PAP-Net (see Section 3.3).

On PASCAL, we compare our method against the state-of-the-art ASTMT [30]. Note that a direct comparison with ASTMT is also not straightforward, as this model is by design single-scale and heavily based on a DeepLab-v3+ (DLv3+) backbone network that contains dilated convolutions. Due to the latter, simply plugging the same DLv3+ backbone into MTI-Net would break the multi-scale features required to uniquely model the task interactions at a multitude of scales. Yet, we provide a fair comparison with ASTMT by combining it with a ResNet-50 FPN backbone, to show that it is not just using a multi-scale backbone that leads to improved results. We omit a comparison with PAD-Net on PASCAL, as we found it to not scale well to the set of tasks in this dataset.

Table 2: Ablation studies on (a) NYUD-v2 and (b) PASCAL using an HRNet-18 backbone network. Auxiliary tasks are indicated between brackets.

(a) Results on NYUD-v2.

Method	Seg ↑	Dep ↓	Δ_m% ↑
Single task	33.18	0.667	+ 0.00
MTL	32.09	0.668	- 1.71
PAD-Net	32.80	0.660	- 0.02
PAD-Net (N)	33.85	0.658	+ 1.65
PAD-Net (N+E)	32.92	0.655	+ 0.52
Ours (w/o FPM)	34.38	0.640	+ 3.85
Ours (w/o FPM) (N)	34.49	0.642	+ 3.84
Ours (w/o FPM) (N+E)	34.68	0.637	+ 4.48
Ours (w/ FPM)	35.12	0.620	+ 6.40
Ours (w/ FPM) (N)	36.22	**0.600**	+ 9.57
Ours (w/ FPM) (N+E)	**37.49**	0.607	**+ 10.91**

(b) Results on PASCAL.

Method	Seg ↑	Parts ↑	Sal ↑	Edge ↑	Norm ↓	Δ_m% ↑
Single task	60.07	60.74	67.18	69.70	14.59	+ 0.00
MTL (s)	54.53	59.54	65.60	-	-	- 4.26
MTL (a)	53.60	58.45	65.13	70.60	15.08	- 3.70
Ours (s)	64.06	62.39	68.09	-	-	+ 3.35
Ours (s)(E)	64.98	62.90	67.84	-	-	+ 3.98
Ours (s)(N)	63.74	61.75	67.90	-	-	+ 2.69
Ours (s)(E+N)	64.33	62.33	68.00	-	-	+ 3.36
Ours (a)	64.27	62.06	68.00	73.40	14.75	+ 2.74

3.2 Ablation studies

Network components. In Table 2 we visualize the results of our ablation studies on NYUD-v2 and PASCAL with an HRNet18 backbone to verify how different components of our model contribute to the multi-task improvements. Additional results using different backbones are in the supplementary materials.

We focus on the smaller NYUD-v2 dataset first (see Table 2a), that contains arguably related tasks. These are semantic segmentation (Seg) and depth prediction (Dep) as main tasks, edge detection (E) and surface normals (N) as auxiliary tasks. The MTL baseline (i.e. a shared encoder with task-specific heads) has lower performance (−1.71%) than the single-task models. This is inline with prior work [45,30]. PAD-Net retains performance over the set of single-task models (−0.02%), and improves when adding the auxiliary tasks (+0.52%). Using our model without the FPM between scales further improves the results (w/o auxiliary tasks: +3.85%, w/ auxiliary tasks: +4.48%). When including the FPM another significant boost in performance is achieved (+6.40%). Further adding the auxiliary tasks can help to improve the quality of our predictions (+10.91%).

Table 2b shows the ablation on PASCAL. We discriminate between a *small set (s)* and a *complete set (a)* of tasks. The small set contains the high-level (semantic and human parts segmentation) and mid-level (saliency) vision tasks. The complete set also adds the low-level (edges and normals) vision tasks. The MTL baseline leads to decreased performance, −4.26% and −3.70% on the small and complete set respectively. Instead, our model improves over the set of single-task models (+3.35%) on the small task set (s), where we obtain solid improvements on all tasks. We also report the influence of adding additional auxiliary tasks to the front-end of the network. Adding edges improves the multi-task performance to 3.98%, adding normals slightly decreases it to +2.69%, while adding both keeps it stable +3.36%. Finally, when learning all five tasks together, our model outperforms (+2.74%) the set of single-task models. In general, all tasks gain significantly, except for normals, where we observe a small decrease in performance. We argue that this is due to the inevitable negative transfer that characterizes all models with shared operations (also [48,30,52]). Yet, to the best of our knowledge, this is the first work to not only report overall improved

Table 3: Influence of using a different number of scales for the backbone network on NYUD-v2.

Method	Seg ↑	Dep ↓	Δ_m% ↑
ST	33.18	0.667	+ 0.00
1/4 (Pad-Net)	32.80	0.660	- 0.02
1/4, 1/8	34.88	0.650	+ 3.80
1/4, 1/8, 1/16	35.01	0.630	+ 5.53
1/4, 1/8, 1/16, 1/32 (Ours)	35.12	0.620	+ 6.40

Table 4: Ablating the information flow within the proposed MTI-Net model on NYUD-v2.

Method	Seg ↑	Dep ↓	Δ_m% ↑
ST	33.18	0.667	+ 0.00
Front-end @ 1/32 scale	32.02	0.670	- 1.87
Front-end @ 1/16 scale	33.02	0.660	+ 0.02
Front-end @ 1/8 scale	33.67	0.640	+ 2.72
Front-end @ 1/4 scale	34.05	0.633	+ 3.78
Final output	35.12	0.620	+ 6.40

Table 5: Comparison with the state-of-the-art on PASCAL.

Model	Backbone	Seg ↑		Parts ↑		Sal ↑		Edge ↑		Norm ↓		Δ_m ↑ (ST)	Δ_m ↑ (R50-FPN)
		ST	MT	ST	MT	ST	MT	ST	MT	ST	MT		
ASTMT [30]	R26-DLv3+	64.9	64.6	57.1	57.3	64.2	64.7	71.3	71.0	14.9	15.0	- 0.11	- 3.42
	R50-DLv3+	68.3	68.0	60.70	61.1	65.4	65.7	72.7	72.4	14.6	14.7	- 0.04	- 0.08
	R50-FPN	67.7	66.8	61.8	61.1	67.2	66.1	71.1	70.9	14.8	14.7	- 0.87	- 0.87
Ours	R18-FPN	64.5	65.7	57.4	61.6	66.4	66.8	68.2	73.9	14.8	14.6	+ 3.84	+ 0.29
	R50-FPN	67.7	66.6	61.8	63.3	67.2	66.6	71.1	74.9	14.8	14.6	+ 1.36	+ 1.36
	HRNet-18	60.1	64.3	60.7	62.1	67.2	68.0	69.7	73.4	14.6	14.8	+ 2.74	- 0.02

multi-task performance, but also to maximize the gains over the single-task models, when jointly predicting an increasing and diverse set of tasks. We refer to Figure 4 for qualitative results obtained with an HRNet-18 backbone.

Influence of scales. So far, our experiments included all four scales of the backbone network (1/4, 1/8, 1/16, 1/32). Here, we study the influence of using a different number of scales for the backbone. Table 3 summarizes this ablation on NYUD-v2. Note that the use of a single scale (1/4) reduces our model to a PAD-Net like architecture. Using an increasing number of scales (1/4 vs + 1/8 vs + 1/16, ...) gradually improves performance. The results confirm our hypothesis from Section 2.2, i.e. task interactions should be modeled at multiple scales.

Information flow. To quantify the flow of information, we measure the performance of the initial task predictions at different locations in the front-end of the network. Table 4 illustrates the results on NYUD-v2. We observe that the performance gradually increases at the higher scales, due to the information that is being propagated from the lower scales via the FPM. The final prediction after aggregating the information from all scales is further improved substantially.

3.3 Comparison with the state-of-the-art

Comparison on PASCAL. Table 5 visualizes the comparison of our model against ASTMT on PASCAL. We report the multi-tasking performance both w.r.t. the single-task models using the same backbone (ST) and the single-task models based on the R50-FPN backbone. As explained, in the only possible fair comparison, i.e. when using the same R50-FPN backbone, our model achieves higher multi-tasking performance compared to ASTMT (+1.36% vs −0.87%). Yet, as ASTMT is by design single-scale and heavily based on DLv3+, we also report results using different backbones. Overall, MTI-Net achieves significantly higher gains over its single-task variants compared to ASTMT (see $\Delta_m \uparrow$ (ST)). Surprisingly, we find that our model with R18-FPN backbone even outperforms

Table 6: Comparison with the state-of-the-art on NYUD-v2.

(a) Results on depth estimation.

Method	rmse	rel	δ_1	δ_2	δ_3
HCRF [21]	0.821	0.232	0.621	0.886	0.968
DCNF [25]	0.824	0.230	0.614	0.883	0.971
Wang [47]	0.745	0.220	0.605	0.890	0.970
NR forest [37]	0.774	0.187	-	-	-
Xu [49]	0.593	0.125	0.806	0.952	0.986
PAD-Net [48]	0.582	**0.120**	0.817	0.954	0.987
PAP-Net [52]	0.530	0.144	0.815	0.962	0.992
ST - HRNet48-V2	0.547	0.138	0.828	0.966	0.993
Ours - HRNet48-V2	**0.529**	0.138	**0.830**	**0.969**	**0.993**

(b) Results on semantic segmentation.

Method	pixel-acc	mean-acc	IoU
FCN [27]	60.0	49.2	29.2
Context [23]	70.0	53.6	40.6
Eigen [8]	65.6	45.1	34.1
B-SegNet [16]	68.0	45.8	32.4
RefineNet-101 [22]	72.8	57.8	44.9
PAD-Net [48]	75.2	62.3	50.2
TRL-ResNet50 [51]	76.2	56.3	46.4
PAP-Net [52]	**76.2**	62.5	**50.4**
ST - HRNet48-V2	73.4	58.1	45.7
Ours - HRNet48-V2	75.3	**62.9**	49.0

Table 7: Computational resource analysis (number of parameters and FLOPS). The results are reported relative to the single-task models.

(a) Results on NYUD-v2 using an HRNet-18 backbone.

Method	Params (M)	FLOPS (G)	$\Delta_m\%$
Single Task	8.0	22.0	+0.00%
Multi-Task	−50%	−45%	−1.71%
PAD-Net	−15%	+204%	−0.02%
MTI-Net (Ours)	+57%	−13%	+6.40%

(b) Results on PASCAL using a ResNet-50 FPN backbone.

Method	Params (M)	FLOPS (G)	$\Delta_m\%$
Single Task	140	219	+0.00%
Multi-Task	−75.0%	−66%	−4.55%
ASTMT	−51.0%	−1.0%	−0.87%
Ours	−35.0%	−19.9%	+1.36%

the deeper ASTMT R50-DLv3+ model in terms of multi-tasking performance (+0.29% vs −0.08%), despite the fact that the ASTMT single-task models perform better than ours, due to the use of the stronger DLv3+ backbone. Note that we are the first to report consistent multi-task improvements when solving such a diverse task dictionary.

Comparison on NYUD-v2. Table 6 shows a comparison with the state-of-the-art approaches on NYUD-v2. We leave out methods that rely on extra input modalities, or additional training data. As these methods are built on top of stronger single-scale backbones, we also use the multi-scale HRNet48-v2 backbone here. Again, our model improves w.r.t the single-task models. Furthermore, we perform on par with the state-of-the-art on the depth estimation task, while performing slightly worse on the semantic segmentation task. We refer the reader to the supplementary materials for qualitative results.

Resource analysis. We compare our model in terms of computational requirements (number of parameters and FLOPS) against PAD-Net and ASTMT. The comparison with PAD-Net is performed on NYUD-v2 using the HRNet-18 backbone, while for the comparison with ASTMT on PASCAL we use a ResNet-50 FPN backbone. Results for every method relative to the single-tasking models are reported in Table 7.

On NYUD-v2, MTI-Net reduces the number of FLOPS while improving the performance compared to the single-task models. The reason for the increased amount of parameters is the use of a shallow backbone, and the small number of tasks (i.e. 2). Furthermore, we significantly outperform PAD-Net in terms of FLOPS and multi-task performance. This is due to the fact that PAD-Net

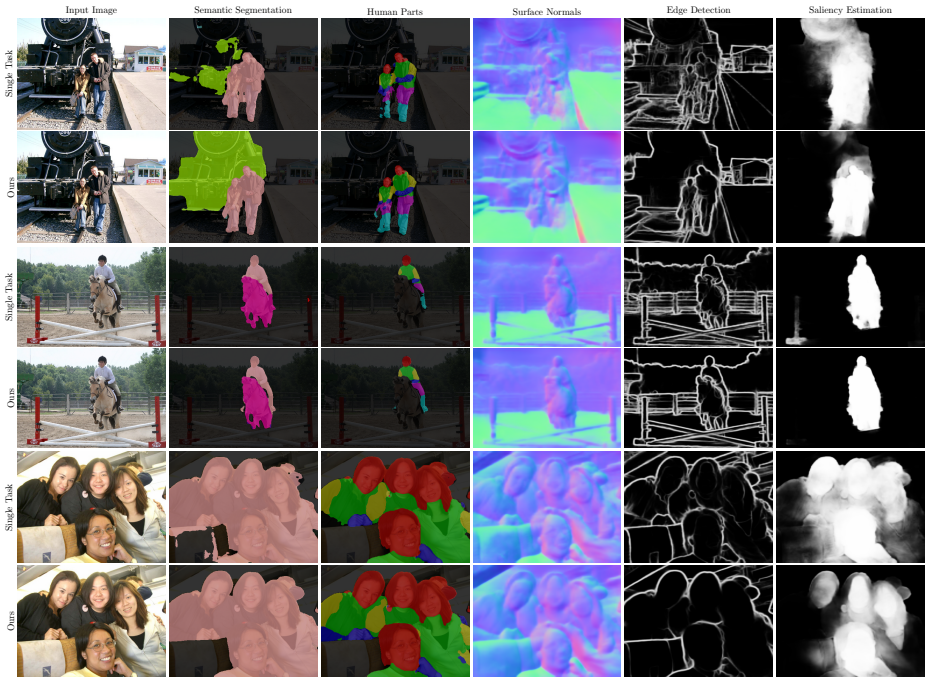


Fig. 4: **Qualitative results on PASCAL.** We compare the predictions made by a set of single-task models (first row for every image) against the predictions made by our MTI-Net (second row for every image). Differences can be seen for semantic segmentation, edge detection and saliency estimation.

performs the multi-modal distillation at a single higher scale (1/4) with $4 \cdot C$ channels, C being the number of backbone channels at a single scale. Instead, we perform most of the computations at smaller scales (1/32, 1/16, 1/8), while operating on only C channels at the higher scale (1/4).

On PASCAL, we significantly improve on all three metrics compared to the single-task models. We also outperform ASTMT in terms of FLOPS and multi-task performance, as the latter has to perform a separate forward pass per task.

4 Conclusion

We have shown the importance of modeling task interactions at multiple scales, enabling tasks to maximally benefit each other. We achieved this by introducing dedicated modules on top of an off-the-shelf multi-scale feature extractor, i.e. multi-scale multi-modal distillation, feature propagation across scales, and feature aggregation. Our multi-task model delivers on the full potential of multi-task learning, i.e. smaller memory footprint, reduced number of calculations and better performance. Our experiments show that our multi-task models consistently outperform their single-tasking counterparts by medium to large margins.

Acknowledgment This work is sponsored by the Flemish Government under the Artificiele Intelligentie (AI) Vlaanderen programme. The authors also acknowledge support by Toyota via the TRACE project and MACCHINA (KU Leuven, C14/18/065).

References

1. Bansal, A., Chen, X., Russell, B., Gupta, A., Ramanan, D.: Pixelnet: Representation of the pixels, by the pixels, and for the pixels. arXiv preprint arXiv:1702.06506 (2017)
2. Caruana, R.: Multitask learning. Machine learning **28**(1), 41–75 (1997)
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 801–818 (2018)
4. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.: Detect what you can: Detecting and representing objects using holistic models and body parts. In: CVPR. pp. 1971–1978 (2014)
5. Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A.: Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: ICML (2018)
6. Doersch, C., Zisserman, A.: Multi-task self-supervised visual learning. In: ICCV. pp. 2051–2060 (2017)
7. Dvornik, N., Shmelkov, K., Mairal, J., Schmid, C.: Blitznet: A real-time deep network for scene understanding. In: ICCV. pp. 4154–4162 (2017)
8. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV. pp. 2650–2658 (2015)
9. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88**(2), 303–338 (2010)
10. Girshick, R.: Fast r-cnn. In: ICCV. pp. 1440–1448 (2015)
11. Guo, M., Haque, A., Huang, D.A., Yeung, S., Fei-Fei, L.: Dynamic task prioritization for multitask learning. In: ECCV (2018)
12. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: ECCV. pp. 345–360. Springer (2014)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2961–2969 (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
15. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141 (2018)
16. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint arXiv:1511.02680 (2015)
17. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR (2018)
18. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: CVPR. pp. 6399–6408 (2019)
19. Kokkinos, I.: Pushing the boundaries of boundary detection using deep learning. arXiv preprint arXiv:1511.07386 (2015)

20. Kokkinos, I.: Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In: CVPR (2017)
21. Li, B., Shen, C., Dai, Y., Van Den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: CVPR. pp. 1119–1127 (2015)
22. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: CVPR. pp. 1925–1934 (2017)
23. Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: CVPR. pp. 3194–3203 (2016)
24. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)
25. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. TPAMI **38**(10), 2024–2039 (2015)
26. Liu, S., Johns, E., Davison, A.J.: End-to-end multi-task learning with attention. In: CVPR (2019)
27. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
28. Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., Feris, R.: Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In: CVPR (2017)
29. Maninis, K.K., Pont-Tuset, J., Arbeláez, P., Van Gool, L.: Convolutional oriented boundaries: From image segmentation to high-level tasks. TPAMI **40**(4), 819–833 (2017)
30. Maninis, K.K., Radosavovic, I., Kokkinos, I.: Attentive single-tasking of multiple tasks. In: CVPR. pp. 1851–1860 (2019)
31. Martin, D.R., Fowlkes, C.C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. TPAMI (5), 530–549 (2004)
32. Misra, I., Shrivastava, A., Gupta, A., Hebert, M.: Cross-stitch networks for multi-task learning. In: CVPR (2016)
33. Neven, D., De Brabandere, B., Georgoulis, S., Proesmans, M., Van Gool, L.: Fast scene understanding for autonomous driving. In: IV Workshops (2017)
34. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV. pp. 483–499. Springer (2016)
35. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015)
36. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
37. Roy, A., Todorovic, S.: Monocular depth estimation using neural regression forest. In: CVPR. pp. 5506–5514 (2016)
38. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
39. Sener, O., Koltun, V.: Multi-task learning as multi-objective optimization. In: NIPS (2018)
40. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
41. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV. pp. 746–760. Springer (2012)

42. Sinha, A., Chen, Z., Badrinarayanan, V., Rabinovich, A.: Gradient adversarial training of neural networks. arXiv preprint arXiv:1806.08028 (2018)
43. Standley, T., Zamir, A.R., Chen, D., Guibas, L., Malik, J., Savarese, S.: Which tasks should be learned together in multi-task learning? arXiv preprint arXiv:1905.07553 (2019)
44. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR. pp. 5693–5703 (2019)
45. Vandenhende, S., Georgoulis, S., De Brabandere, B., Van Gool, L.: Branched multi-task networks: Deciding what layers to share. arXiv preprint arXiv:1904.02920 (2019)
46. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. arXiv preprint arXiv:1908.07919 (2019)
47. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L.: Towards unified depth and semantic prediction from a single image. In: CVPR. pp. 2800–2809 (2015)
48. Xu, D., Ouyang, W., Wang, X., Sebe, N.: Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In: CVPR. pp. 675–684 (2018)
49. Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., Ricci, E.: Structured attention guided convolutional neural fields for monocular depth estimation. In: CVPR. pp. 3917–3925 (2018)
50. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: CVPR (2018)
51. Zhang, Z., Cui, Z., Xu, C., Jie, Z., Li, X., Yang, J.: Joint task-recursive learning for semantic segmentation and depth estimation. In: ECCV. pp. 235–251 (2018)
52. Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., Yang, J.: Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In: CVPR. pp. 4106–4115 (2019)
53. Zhao, X., Li, H., Shen, X., Liang, X., Wu, Y.: A modulation module for multi-task learning with applications in image retrieval. In: ECCV (2018)

Table S1: Multi-task learning on PASCAL using a ResNet-18 FPN backbone.

Method	Seg ↑	Parts ↑	Sal ↑	Edge ↑	Norm ↓	Δ_m ↑
Single task	64.49	57.43	66.38	68.20	14.77	+ 0.00
MTL (s)	54.51	55.12	64.76	-	-	- 7.32
MTL (a)	59.61	56.88	64.96	70.60	15.17	- 1.80
Ours (s)	65.47	61.32	66.37	-	-	+ 2.77
Ours (s)(E)	65.93	62.21	66.80	-	-	+ 3.61
Ours (s)(N)	64.99	61.09	66.80	-	-	+ 2.52
Ours (s)(E+N)	65.46	61.71	66.62	-	-	+ 3.06
Ours (a)	65.69	61.59	66.76	73.90	14.55	+ 3.84

A Training setup

We include additional details of the training setup used for each experiment. We considered two different multi-scale backbone networks, i.e. HRNet and FPN. For HRNet, we use bilinear upsampling and concatenation followed by two convolutional layers to decode the multi-scale features in the feature aggregation unit. For FPN, the feature aggregation module decodes the multi-scale features as in panoptic feature pyramid networks [18]. In both cases, the non-linear function that produces the task attention mask in the FPM is implemented as two basic residual blocks – that aggressively reduce the number of channels – followed by a 1×1 convolutional layer.

A.1 NYUD-v2

We applied the data augmentation strategy of PAD-Net [48]. The RGB and depth images were randomly scaled with the selected ratio in $\{1, 1.2, 1.5\}$, and randomly horizontally flipped. The model was trained for 80 epochs with an Adam optimizer with initial learning rate 1e-4 and batches of size 6. We used a poly learning rate decay scheme.

A.2 PASCAL

We essentially plugged our model into the code base that was shared by [30]. In particular, the single-task models were trained with stochastic gradient descent with momentum 0.9. We used batches of size 8 and a poly learning rate decay scheme. The initial learning rate was 0.01. We applied weight decay $\lambda = 1e-4$. These hyperparameters are the same as the ones used in [30], ensuring fair comparison. The multi-task baseline models were trained using the same hyperparameters. The multi-task loss weighing was taken from [30]. We also tested the use of an Adam optimizer, but this did not yield better results.

Our MTI-Net was trained under the same settings as the single-task models, but we used an Adam optimizer with initial learning rate 1e-4. We re-used the loss weights from before to weight the losses from the initial task predictions.

B Extra experiments on PASCAL

We perform an additional ablation experiment using a ResNet-18 FPN backbone in Table S1. The conclusions are similar to the ones reported for the HRNet-18 in Table 2b of the main paper. This shows that our method can be used in combination with various backbone architectures. Again, our model improves over the single-tasking models, both for the small (+2.77%) and complete (+3.84%) set of tasks. We also consider the effect of adding additional auxiliary tasks when predicting the small task set. When adding edge detection as an auxiliary task, the results are further improved (+2.77% to +3.61%). However, this is not the case when we add surface normals prediction as an auxiliary task (+2.77% to 2.52%). We observe a similar effect when including both edge detection and surface normals prediction (3.61% to 3.06%). We believe that this is due to the approximate nature of the surface normals in this dataset, as the latter were obtained through distillation, and as such they are rather noisy.

C Extra experiments on NYUD-v2

This section contains additional results on the NYUD-v2 dataset. Section C.1 gives a more detailed view on the ablation studies that we performed on NYUD-v2 using an HRNet-18 backbone. Note that the main results of this experiment were already discussed in the experiments section of the paper. In Section C.2, we perform an additional experiment using an FPN backbone based on ResNet-18.

C.1 HRNet18-V2

Table S2 contains additional metrics for the depth estimation and semantic segmentation task on the NYUD-v2 dataset, when using an HRNet-18 backbone. This is an extension to the metrics shown in Table 2a of the paper.

C.2 FPN - ResNet-18

We repeated a smaller version of our ablation studies on the NYUD-v2 dataset when using an FPN backbone based on ResNet-18. Table S3 contains the results. We end up at similar findings compared to the model based on HRNet-18. Again, we see a significant improvement over the set of single-task models. Additionally, we find that the use of auxiliary tasks can help to improve the quality of the predictions, as the latter are not distilled in this case.

D Qualitative results on NYUD-v2

Figure S1 shows predictions made by our HRNet-48 model on images from the NYUD-v2 test set. The quantitative results were already reported in Table 6 of the paper.

Table S2: Ablation studies on NYUD-v2 using an HRNet18-V2 backbone. Auxiliary tasks are indicated in brackets.

(a) Results on the depth estimation task.

Method	rmse ↓	rel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Single task	0.667	0.186	0.731	0.931	0.981
MTL	0.668	0.193	0.717	0.927	0.980
PAD-Net	0.660	0.189	0.726	0.930	0.981
PAD-Net (N)	0.658	0.187	0.726	0.932	0.982
PAD-Net (N+E)	0.655	0.184	0.731	0.934	0.982
Ours - w/o FPM	0.640	0.181	0.747	0.937	0.982
Ours - w/o FPM (N)	0.642	0.175	0.753	0.940	0.983
Ours - w/o FPM (N+E)	0.637	0.174	0.757	0.939	0.984
Ours - w/ FPM	0.620	0.161	0.781	0.946	0.986
Ours - w/ FPM (N)	0.600	0.162	0.788	0.947	0.985
Ours - w/ FPM (N+E)	0.607	0.166	0.783	0.945	0.985

(b) Results on the semantic segmentation task.

Method	pixel-acc ↑	mean-acc ↑	IoU ↑
Single task	65.04	45.07	33.18
MTL	64.61	43.55	32.09
PAD-Net	65.00	44.61	32.80
PAD-Net (N)	64.77	46.28	33.85
PAD-Net (N+E)	65.05	44.79	32.92
Ours - w/o FPM	65.52	45.98	34.38
Ours - w/o FPM (N)	65.27	46.63	34.49
Ours - w/o FPM (N+E)	66.15	46.97	34.68
Ours - w/ FPM	66.30	47.85	35.12
Ours - w/ FPM (N)	66.98	49.04	36.22
Ours - w/ FPM (N+E)	68.03	51.05	37.49

Table S3: Additional results on NYUD-v2 when using an FPN backbone based on ResNet-18. Similarly to Table 2, auxiliary tasks are indicated in brackets.

(a) Multi-task learning performance.

Method	Seg (IoU) ↑	Dep (rmse) ↓	$\Delta_m\%$
Single task	34.46	0.659	+0.00
MTL	33.52	0.665	-1.82
PAD-Net	34.15	0.662	-0.69
PAD-Net (N)	34.18	0.657	-0.23
PAD-Net (N+E)	34.60	0.668	-0.45
Ours	36.01	0.630	+4.43
Ours (N)	36.81	0.628	+5.74
Ours (N+E)	36.65	0.618	+6.27

(b) Results on the depth estimation task.

Method	rmse ↓	rel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Single task	0.659	0.183	0.730	0.935	0.982
MTL	0.665	0.190	0.726	0.930	0.980
PAD-Net	0.662	0.188	0.731	0.931	0.979
PAD-Net (N)	0.657	0.185	0.735	0.934	0.980
PAD-Net (N+E)	0.668	0.185	0.729	0.933	0.980
Ours	0.630	0.173	0.767	0.939	0.981
Ours (N)	0.628	0.180	0.755	0.939	0.982
Ours (N+E)	0.618	0.169	0.768	0.944	0.984

(c) Results on the semantic segmentation task.

Method	pixel-acc ↑	mean-acc ↑	IoU ↑
Single task	65.51	46.50	34.46
MTL	64.85	45.33	33.52
PAD-Net	65.23	45.65	34.15
PAD-Net (N)	65.07	45.80	34.18
PAD-Net (N+E)	65.68	46.77	34.60
Ours	66.44	49.03	36.01
Ours (N)	66.89	50.50	36.81
Ours (N+E)	67.23	49.93	36.65

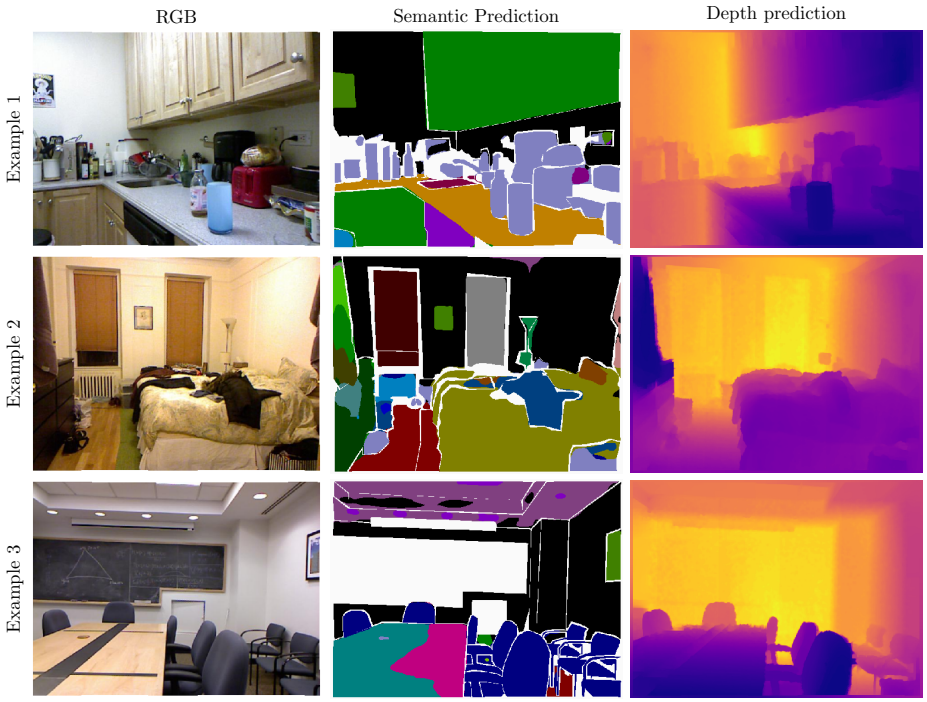


Fig. S1: **Qualitative results on NYUD-v2:** Semantic and depth predictions made by our HRNet-48 model.