

Automated Search for Resource-Efficient Branched Multi-Task Networks

David Bruggemann

brdavid@vision.ee.ethz.ch

Menelaos Kanakis

kanakism@vision.ee.ethz.ch

Stamatios Georgoulis

georgous@vision.ee.ethz.ch

Luc Van Gool

vangool@vision.ee.ethz.ch

Computer Vision Lab

ETH Zurich

Switzerland

Abstract

The multi-modal nature of many vision problems calls for neural network architectures that can perform multiple tasks concurrently. Typically, such architectures have been handcrafted in the literature. However, given the size and complexity of the problem, this manual architecture exploration likely exceeds human design abilities. In this paper, we propose a principled approach, rooted in differentiable neural architecture search, to automatically define branching (tree-like) structures in the encoding stage of a multi-task neural network. To allow flexibility within resource-constrained environments, we introduce a proxyless, resource-aware loss that dynamically controls the model size. Evaluations across a variety of dense prediction tasks show that our approach consistently finds high-performing branching structures within limited resource budgets.

1 Introduction

Over the last decade neural networks have shown impressive results for a multitude of tasks. This is typically achieved by designing network architectures that satisfy the specific needs of the task at hand, and training them in a fully supervised fashion [1, 2, 3]. This well-established strategy assumes that each task is tackled in isolation, and consequently, a dedicated network should be learned for every task. However, real-world problems are inherently multi-modal (e.g., an autonomous car should be able to detect road lanes, recognize pedestrians in its vicinity, semantically understand its surroundings, estimate its distance from other objects, etc.), which calls for architectures that can perform multiple tasks simultaneously.

Motivated by these observations researchers started designing multi-task architectures that, given an input image, can produce predictions for all desired tasks [4]. Arguably the most characteristic example is branched multi-task networks [5, 6, 7, 8, 9], where a shared encoding stage branches out to a set of task-specific heads that decode the shared features to yield task predictions. Follow-up works [10, 11, 12, 13, 14, 15] proposed more

advanced mechanisms for multi-task learning, however, despite their advantages, they similarly assume that the multi-task architecture can be handcrafted prior to training. In a different vein, some works [25, 24] opted for a semi-automated architecture design where the branching occurs at finer locations in the encoding stage. Task groupings at each branching location are determined based on a measure of ‘task relatedness’ in pre-trained networks. However, defining branching points based on such offline criteria disregards potential optimization benefits resulting from jointly learning particular groups of tasks [10].

Generally, finding an architecture suitable for multi-task learning poses great challenges in trying to balance conflicting objectives. On the one hand, a multi-task network should perform comparably to its single-task counterparts. This is not trivial as task interference¹ can significantly affect individual task performance. On the other hand, unlike [9, 22, 29], a multi-task network should keep within a low computational budget during inference with respect to the single-task case. The relative importance of task performance vs. computational efficiency depends on the application, highlighting the necessity of being able to adapt the architecture design flexibly. This calls for automatic architecture search techniques to mitigate the effort accompanying architecture handcrafting.

In this paper, we address the aforementioned needs and propose Branched Multi-Task Architecture Search (BMTAS²), a principled approach to automatically determine encoder branching in multi-task architectures. To avoid a brute-force or greedy search as in [24] and [25] respectively, we build a differentiable neural architecture search algorithm with a search space directly encompassing all possible branching structures. Our approach is end-to-end trainable, and allows flexibility in the model construction through the introduction of a proxyless, resource-aware loss. Experimental analysis on multiple dense prediction tasks shows that models generated by our method effectively balance the trade-off between overall performance and computational cost.

2 Related Work

Multi-Task Learning (MTL) in deep neural networks addresses the problem of training a single model that can perform multiple tasks concurrently. To this end, a first group of works [9, 22, 24, 38] incorporated feature sharing mechanisms on top of a set of task-specific networks. Typically, these *soft parameter sharing* approaches scale poorly to an increasing number of tasks, as the parameter count usually exceeds the single-task case. A second group of works [0, 17, 19, 24, 31] shared the majority of network operations, before branching out to a set of task-specific heads that produce the task predictions. In these *hard parameter sharing* approaches the branching point is manually determined prior to training, which can lead to task interference if ‘unrelated’ (groups of) tasks are selected.

Follow-up works employed different techniques to address the MTL problem. PAD-Net [38] and MTI-Net [24] proposed to refine the initial task predictions by distilling cross-task information. Dynamic task prioritization [17] constructed a hierarchical network that dynamically prioritizes the learning of ‘easy’ tasks at earlier layers. ASTMT [27] and RCM [16] adopted a task-conditional approach, where only one task is forward-propagated at a time, to sequentially generate all task predictions. All the above mentioned approaches assume that the multi-task architecture can be handcrafted. However, given the size and

¹Task interference is a well-documented problem in multi-task networks [16, 19, 27, 30] where important information for one task might be a nuisance for another, leading to conflicts in the optimization of shared parameters.

²Reference code at <https://github.com/brdav/bmtas>

complexity of the problem, this manual exploration likely exceeds human design abilities. In contrast, stochastic filter groups [4] re-purposed the convolution kernels in each layer to support shared or task-specific behavior, but their method only operates at the channel level of the network.

Our work is closer to [25, 42] where branching structures are generated in the encoding stage of the network. However, instead of determining the task groupings at each branching location in pre-trained networks via a brute-force or greedy algorithm, we propose a principled solution rooted in differentiable neural architecture search that is trainable end-to-end.

Neural Architecture Search (NAS) aims to automate the procedure of designing neural network architectures in contrast to the established protocol of manually drafting them based on prior knowledge. To achieve this, a first group of works [53, 49, 50] used a recurrent neural network to sample architectures from a pre-defined search space and trained it with reinforcement learning to maximize the expected accuracy on a validation set. A second group of works [34, 35] employed evolutionary algorithms to gradually evolve a population of models through mutations. Despite their great success, these works are computationally demanding during the search phase, which motivated researchers to explore differentiable NAS [21] through continuous relaxation of the architecture representation. SNAS [47] built upon Gumbel-Softmax [15, 26] to propose an effective search gradient. Follow-up works [8, 45] introduced resource constraints into the optimization pipeline to control the model size.

In general, NAS works have mainly focused on the image classification task, with a few exceptions that addressed other tasks too (e.g., semantic segmentation [20]). Moreover, each task is tackled in isolation, i.e., a dedicated architecture is generated for each task. When it comes to MTL however, there is a need to design architectures that perform well across a variety of tasks. Routing networks [36] made a first attempt in this direction by determining the connectivity of a network’s function blocks through routing, but they focus on classification tasks, as opposed to the more challenging dense prediction considered in this paper. MTL-NAS [10] extended NDDR-CNN [9] to automatically find the feature fusion mechanisms on top of the task-specific networks, but their search is solely limited to the feature sharing mechanisms. In contrast, we automatically generate branching (or tree-like) structures in the encoding stage of the network, which allows more explicit control over the size of the final model.

3 Method

Given a set of dense prediction tasks and an arbitrary neural architecture, our goal is to find resource-efficient branching structures in the encoder, that promote the sharing of general-purpose features between tasks and the decoupling of task-specific features across tasks. In this section, we elaborate on three key components of the proposed BMTAS, which builds upon differentiable NAS to tackle the aforementioned objective: the structure of the search space (Section 3.1), the algorithm to traverse that search space (Section 3.2), and a novel objective function to enforce resource efficiency (Section 3.3).

3.1 Search Space

In contrast to most established NAS works [21, 53, 49, 50] which only search for subcomponents (e.g., a cell) of the target architecture, our search space directly encompasses all possible branching structures for a given number of T tasks in an encoder network. As

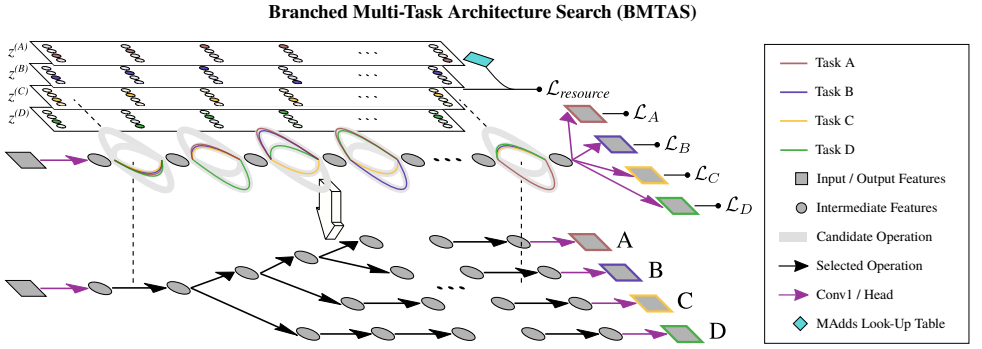


Figure 1: Visualization of the construction of a branched architecture (bottom) from a supergraph (top). For each of the tasks A to D, a subgraph is sampled from the supergraph (colored edges), using the learnable masks $z^{(\cdot)}$. To form a branching structure, the subgraphs are combined according to the present task groupings. During the architecture search, the masks $z^{(\cdot)}$ are learned by minimizing a resource loss $\mathcal{L}_{resource}$ (computed using a look-up table) and task performance losses \mathcal{L}_A to \mathcal{L}_D simultaneously.

shown in Figure 1, we can describe our search space with a directed acyclic graph, where the vertices represent intermediate feature tensors and the edges operations (e.g., bottleneck blocks in ResNet-50) of the chosen network. For an encoder with L layers the graph has length L in total, and width T between consecutive vertices. Parallel edges denote candidate operations, which are (non-parameter-sharing) duplicates of the original operation in the respective layer. The operation parameters of each duplicate block are ‘warmed up’ for a few iterations on the corresponding task before the architecture search. Through this procedure, each block is softly assigned to a task. For any one task t , a specific routing (subgraph) through the supergraph can be obtained by sampling operations with a mask $z^{(t)} \in \{0, 1\}^{L \times T}$ with one-hot rows. Any possible branching structure for a set of tasks can then be produced by combining the task-specific routings. Note that, in order for computation sharing to occur in layer l for two tasks, their sampled edges need to coincide in all layers 1 to l . We present a loss function which incentivizes such computation sharing in Section 3.3.

3.2 Search Algorithm

As explained in the previous section, the goal of BMTAS is to find a set of masks $Z = \{z^{(t)} | t \in \{1, \dots, T\}\}$ to sample T subgraphs from the supergraph. We model this set with a parameterized distribution $p_\alpha(Z)$, where $\alpha^{(t)} \in \mathbb{R}^{L \times T}$ represents the unnormalized log probabilities for sampling individual operations for task t . Therefore, the overall optimization problem we want to solve is:

$$\min_{\alpha, \theta} \mathbb{E}_{Z \sim p_\alpha(Z)} [\mathcal{L}(Z, \theta)] \quad (1)$$

θ denotes the neural operation parameters of the supergraph and \mathcal{L} an appropriate loss function (see Section 3.3). By solving this optimization problem, we are maximizing the expected performance (in accordance with the chosen \mathcal{L}) of branching structures sampled from $p_\alpha(Z)$.

Following [47], we can relax the discrete architecture distribution $p_\alpha(Z)$ to be continuous

and differentiable using the gradient estimator proposed in [15, 26]:

$$z_{l,j}^{(t)} = \frac{\exp\left(\left(\alpha_{l,j}^{(t)} + g_{l,j}^{(t)}\right) / \tau\right)}{\sum_{i=1}^T \exp\left(\left(\alpha_{l,i}^{(t)} + g_{l,i}^{(t)}\right) / \tau\right)}, \quad j = 1, \dots, T \quad (2)$$

$g_{l,j}^{(t)} \sim \text{Gumbel}(0, 1)$ is random noise and $\tau > 0$ a temperature parameter. If τ is very large, the sampling distribution is nearly uniform, regardless of the $\alpha^{(t)}$ values. This has the advantage that the resulting gradients in the backpropagation are smooth. For τ close to 0, we approach sampling from the categorical distribution $p_\alpha(Z)$. In practice, we gradually anneal $\tau \rightarrow 0$ during training. Samples from the supergraph are obtained by multiplying the edges in layer l with the softened one-hot vector $z_l^{(t)}$ and summing the output.

The above function, known as Gumbel-Softmax, enables us to directly learn the unnormalized log probability masks $\alpha^{(t)}$ through gradient descent for each task t . During training, we alternate between updating the architecture parameters α and the operation parameters θ , as this empirically leads to more stable convergence in our case. Afterwards, we discretize $\alpha^{(t)}$ using `argmax` to obtain one-hot masks that determine the final routing. As is common practice [21, 47], we retrain the searched architectures from scratch.

3.3 Resource-Aware Objective Function

In the simplest case, the loss function in the overall optimization problem (Equation 1) of BMTAS consists of a weighted sum of the task-specific losses:

$$\mathcal{L}_{\text{tasks}}(Z, \theta) = \sum_{t=1}^T \omega_t \mathcal{L}_t(Z, \theta) \quad (3)$$

Simply searching encoder structures via this objective function yields performance-oriented branching structures, irrespective of the efficiency of the resulting model. We show in Section 4.3 that the outcome resembles separate single-task models, i.e., the tasks stop sharing computations and branch out in early layers, increasing the capacity of the resulting network. To obtain more compact encoder structures and to actively navigate the efficiency vs. performance trade-off, we introduce a resource-aware term $\mathcal{L}_{\text{resource}}$ in the objective function:

$$\mathcal{L}_{\text{search}} = \mathcal{L}_{\text{tasks}} + \lambda \mathcal{L}_{\text{resource}} \quad (4)$$

The emphasis shifts to resource efficiency in the architecture search as λ is increased. We follow related work [21, 30, 44] in choosing the number of multiply-add operations (MAdds) during inference as a surrogate for resource efficiency. The MAdds $C(Z)$ of a sampled branched architecture Z depends on the present *task groupings* in each network layer. We define a task grouping as a *partition* of the set $\{1, \dots, T\}$, where the parts indicate computational sharing. For an encoder with L layers and K possible groupings per layer, the resource objective can be formulated as:

$$\mathbb{E}_{Z \sim p_\alpha(Z)} [C(Z)] = \sum_{l=1}^L \sum_{k=1}^K p_\alpha(\kappa_l = k) c(k, l) \quad (5)$$

where κ_l is the task grouping at layer l , and $c(k, l)$ the MAdds of grouping k at layer l . Note that, $c(k, l)$ can be simply determined by a look-up table, as the computational cost only depends on the layer index and the number of required operations given the grouping.

In contrast to differentiable NAS works [9, 45, 47], we cannot calculate $p_\alpha(\kappa_l = k)$ for each layer independently, since task groupings depend on previous layers. As explained in Section 3.1, two tasks only share computation in layer l provided that they also do so in all preceding layers. Thus, for each task grouping k , we need to map out the set of valid ancestor groupings \mathcal{A}_k . In mathematical terms, \mathcal{A}_k contains every partition (i.e., task grouping) of which k is a *refinement*. Computational sharing with grouping k in layer l only occurs if the groupings in every layer 1 to $(l - 1)$ are in \mathcal{A}_k . Considering this dependency structure, we can decompose $p_\alpha(\kappa_l = k)$ as a recursive formulation of conditional probabilities:

$$p_\alpha(\kappa_l = k) = p_\alpha(\kappa_l = k | \kappa_{l-1}, \dots, \kappa_1 \in \mathcal{A}_k) \sum_{m \in \mathcal{A}_k} p_\alpha(\kappa_{l-1} = m), \quad l = 2, \dots, L \quad (6)$$

The conditional probabilities $p_\alpha(\kappa_l = k | \kappa_{l-1}, \dots, \kappa_1 \in \mathcal{A}_k)$ are independent for each layer and can be easily constructed from the unnormalized log sampling probabilities α .

The derivations presented in this section yield a *proxyless* resource loss function, i.e., it encourages solutions which directly minimize the expected resource cost of the final model. By design, this resource objective function allows us to find tree-like structures without resorting to curriculum learning techniques during the architecture search, which would be infeasible using simpler, indirect constraints (e.g., \mathcal{L}_2 regularization).

4 Experiments

In this section, we first describe the experimental setup (Section 4.1), and consequently evaluate the proposed method on several datasets using different backbones (Section 4.2). We also present ablation studies to validate our search algorithm (Section 4.3), and analyze the resulting task groupings in a case study (Section 4.4).

4.1 Experimental Setup

Datasets. We carry out experiments on PASCAL [8] and NYUD-v2 [44], two popular datasets for dense prediction MTL. For PASCAL we use the PASCAL-Context [8] data split, which includes 4998 training and 5105 testing images, densely labeled for semantic segmentation (SemSeg), human parts segmentation (PartSeg), saliency estimation (Sal), surface normal estimation (Norm), and edge detection (Edge). We adopt the distilled saliency and surface normal labels from [47]. The NYUD-v2 dataset comprises 795 training and 654 testing images of indoor scenes, fully labeled for semantic segmentation (SemSeg), depth estimation (Depth), surface normal estimation (Norm), and edge detection (Edge). All training and evaluation of the baselines and final branched models was conducted on the full-resolution images. Yet, to accelerate the architecture search, we optimize BMTAS using resized input images (1/2 and 2/3 resolution for PASCAL-Context and NYUD-v2 respectively).

Architectures. For all experiments we use a DeepLabv3+ base architecture [8], which was designed for semantic segmentation, but was shown to perform well for various dense prediction tasks [46, 47]. To demonstrate generalization of our method across encoder types, we report results for both MobileNetV2 [49] and ResNet-50 [42] backbones. Note that, for MobileNetV2 we employ a reduced design of the ASPP module (R-ASPP), proposed in [49].

Metrics. We evaluate semantic segmentation, human parts segmentation and saliency estimation using mean intersection over union. For surface normal estimation we use mean error in predicted angles, for edge detection optimal dataset F-measure [48], and for depth

Model	MAdds ↓	SemSeg ↑	PartSeg ↑	Sal ↑	Norm ↓	Edge ↑	Δ_m [%] ↑
Single	15.7B	65.11	57.54	65.41	13.98	69.50	-
Shared	6.0B	59.69	55.96	63.03	16.02	67.80	-6.35
C.-S. [24]	15.7B	63.28	60.21	65.13	14.17	71.10	0.47
NDDR-CNN [9]	21.8B	63.22	56.12	65.16	14.47	68.90	-2.02
MTAN [22]	9.5B	61.55	58.89	64.96	14.74	69.90	-1.73
BMTAS-1	7.5B	61.43	56.77	63.64	14.77	68.20	-3.44
BMTAS-2	9.1B	62.80	57.72	64.92	14.48	68.70	-1.74
BMTAS-3	12.5B	64.07	58.60	64.72	14.27	69.40	-0.60

Table 1: Comparison of BMTAS with MTL baselines on PASCAL-Context using a MobileNetV2 backbone. The resource loss weights λ for BMTAS are chosen on a logarithmic grid: 1e-1, 5e-2 and 2e-2 respectively.

estimation root mean square error. To obtain a single-number performance metric for a multi-task model, we additionally report the average per-task performance drop (Δ_m) with respect to single-task baselines b for model m , as proposed in [22]. It is defined as

$$\Delta_m = \frac{1}{T} \sum_{i=1}^T (-1)^{l_i} (M_{m,i} - M_{b,i}) / M_{b,i} \quad (7)$$

where $l_i = 1$ if a lower value for metric M_i indicates better performance, and $l_i = 0$ otherwise.

Baselines. To eliminate the influence of differences in training setup (data augmentation, hyperparameters, etc.), we compare BMTAS with our own implementation of competing approaches. As a reference, we first report the performance of standard single-task models (‘Single’) and a multi-task model consisting of a fully shared encoder and task-specific heads (‘Shared’). Note that, we consider the ASPP (resp. R-ASPP) module part of the encoder, and as such, share it between all tasks for this model. Furthermore, we create branched encoder structures by two competing methods: *i)* In FAFS [23], a fully shared encoder is first pre-trained and then greedily split layer by layer, starting from the final layer. Branches are separated based on task similarity in sample space, taking into account the complexity of the resulting model. *ii)* In BMN [12], branching structures are found by measuring task affinity with feature map correlations between pre-trained single-task networks. The final models are the result of an exhaustive search over all possible branching configurations, capped at maximum resource cost. Finally, we evaluate the performances of three state-of-the-art MTL approaches: cross-stitch networks (C.-S.) [24], NDDR-CNN [9] and MTAN [22].

4.2 Comparison with the State-of-the-Art

Since the proposed method is targeted towards resource-constrained applications, we optimize two objectives concurrently: multi-task performance and resource cost, and we ultimately seek pareto-efficient solutions. Figure 2 depicts a comparison of the principal encoder branching methods. For each method, three models are generated by varying the complexity penalties in the algorithms. Our models compare favorably to random structures and FAFS, and perform on par with BMN. Unlike BMN however, our approach is end-to-end trainable, and does not rely on an offline brute-force search over the entire space of configurations.

Table 1 and Table 2 show a breakdown of individual task metrics for our method and state-of-the-art MTL approaches on PASCAL-Context and NYUD-v2 using a MobileNetV2

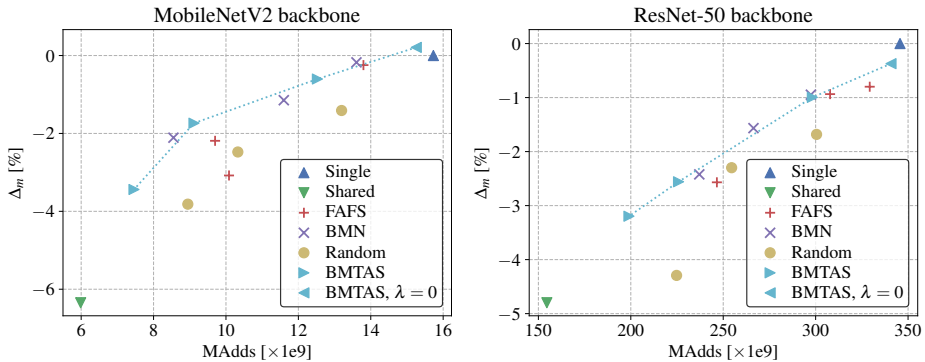


Figure 2: Comparison of multi-task performance (Δ_m , see Section 4.1) as a function of multiply-add operations (MAdds) of encoder branching methods on PASCAL-Context, for a MobileNetV2 backbone (left) and a ResNet-50 backbone (right). For all methods three models with adjustable complexity penalty were evaluated, except ours (BMTAS), for which we additionally report the model obtained by an architecture search without resource loss (no complexity penalty). Note that, BMTAS models favorably balance the efficiency vs. performance trade-off.

and ResNet-50 backbone, respectively. On PASCAL-Context, the BMTAS multi-task performances lie between the ‘Shared’ and ‘Single’ baselines. Although the C.-S. network performs best in this setting, it also requires considerably more MAdds than all the BMTAS models. On the smaller NYUD-v2 datasets, MTL approaches generally perform better compared to the ‘Single’ baseline. Note that, owing to the large number of channels in the feature maps, both NDDR-CNN and MTAN scale inefficiently to a ResNet-50 backbone. The performance boost for these methods therefore comes at the cost of a sizeable increase in resource cost (MAdds).

Overall, BMTAS models gain the advantage in applications where both resource efficiency and task performance are essential. The method is universally applicable to any backbone and thus able to produce compact multi-task models in all tested scenarios. Furthermore, BMTAS is freely adaptable to the application-specific resource budget, a flexibility which is not provided by C.-S., NDDR-CNN or MTAN without changing the backbone. However, the proposed approach also shows some limitations. Notably, the number of possible task groupings grows quickly with increasing tasks³, raising the complexity of computing the resource loss. Combined with the enlarged space of possible branching configurations, this can slow down the architecture search considerably for many-task learning.

For a medium number of tasks however, BMTAS is reasonably efficient. Concretely, the search time was around 1 day on NYUD-v2 (4 tasks, 25000 iterations) and 3.5 days on PASCAL-Context (5 tasks, 50000 iterations) for either backbone using a single V100 GPU.

4.3 Ablation Studies

In Table 3 we present two sets of ablation studies on our search algorithm. As a reference model, we use ‘BMTAS-2’ from Table 1.

³The total number of task groupings for T tasks is given by the Bell number B_T .

Model	MAdds ↓	SemSeg ↑	Depth ↓	Norm ↓	Edge ↑	Δ_m [%] ↑
Single	254.4B	40.08	0.5479	21.67	70.10	-
Shared	122.2B	38.37	0.5766	22.66	70.90	-3.23
C.-S. [14]	254.4B	41.01	0.5380	22.00	70.40	0.77
NDDR-CNN [8]	366.2B	40.88	0.5358	21.86	70.30	0.91
MTAN [12]	600.1B	42.03	0.5191	21.89	70.40	2.38
BMTAS-1	155.8B	40.66	0.5691	21.84	70.70	-0.58
BMTAS-2	223.7B	40.37	0.5413	21.74	69.80	0.30
BMTAS-3	248.3B	41.10	0.5431	21.56	70.10	0.98

Table 2: Comparison of BMTAS with MTL baselines on NYUD-v2 using a ResNet-50 backbone. The resource loss weights λ for BMTAS are 5e-3, 1e-3 and 2e-4 respectively.

Model	SemSeg ↑	PartSeg ↑	Sal ↑	Norm ↓	Edge ↑	Δ_m [%] ↑
BMTAS	62.80	57.72	64.92	14.48	68.70	-1.74
Permutations	63.07±1.09	57.46±0.39	64.25±0.20	14.94±0.25	69.20±0.73	-2.47±0.39
Vanilla	64.29	57.15	64.43	15.06	69.10	-2.34
w/o res. loss	64.70	58.82	65.32	14.14	70.00	0.21
w/o warm-up	62.70	57.93	63.66	14.74	68.60	-2.49
w/o resizing	63.29	57.50	64.06	14.37	68.70	-1.77

Table 3: Ablation studies on PASCAL-Context using a MobileNetV2 backbone. The numbers for ‘Permutations’ reflect the mean and standard deviation of five independent runs.

First, we validate the efficacy of the obtained branching by randomly permuting task-to-branch pairings in the reference structure, before training it (‘Permutations’). We repeat this procedure five times and report the mean scores and standard deviations in Table 3. We also train a simple branching structure consisting of some fully shared layers and simultaneous branching for all tasks (‘Vanilla’). The branching point is chosen such that the number of MAdds coincides with our reference model. Although these two baselines outperform our reference model on SemSeg and Edge, the overall multi-task performance is lower in both cases, suggesting that our method learns to disentangle tasks more efficiently.

Second, we ablate three components of our search algorithm and report the corresponding results in Table 3: *i)* We discard the resource term in the loss (‘w/o res. loss’), i.e., we set $\lambda = 0$ in the total search loss of Equation 4. Both the performance and resource cost (see also Figure 2 left) of the found model are close to the single-task case, as suggested in Section 3.3. *ii)* We disable the task-specific ‘warm-up’ of the parallel candidate operations (‘w/o warm-up’, see Section 3.1) and directly search architectures with ImageNet initialization for θ . The resulting model underperforms our reference model, indicating the importance of softly assigning each candidate operation to a particular task before starting the search: This helps counteract the bias of the algorithm toward full operation sharing early in the training, as well as breaks parameter initialization symmetry. Note that, only the neural operation parameters θ are ‘warmed up’ before the search, while the architecture distribution parameters α are simply initialized with zeros. *iii)* We conduct the architecture search on the full resolution input images (‘w/o resizing’), instead of resizing them as mentioned in Section 4.1. The performance remains approximately equal to the reference model, justifying the use of resized images during architecture search.

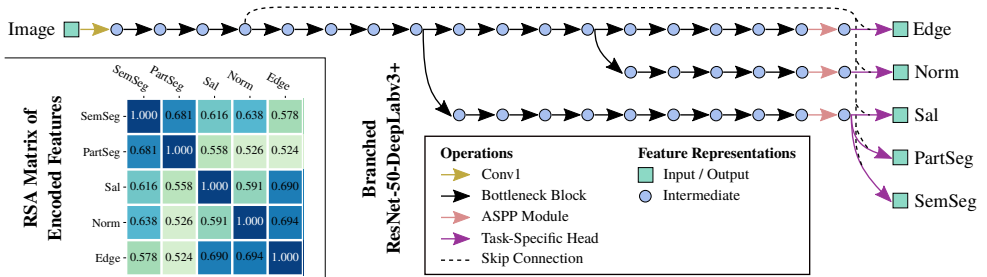


Figure 3: Graph of a ResNet-50-DeepLabv3+ branching structure obtained with our method on the PASCAL-Context dataset (‘BMTAS-2’ in supplementary material Table S2). Edges of the graph indicate operations and vertices indicate feature tensors. On the left, a Representational Similarity Analysis (RSA) matrix determined from correlating encoder output feature maps of the single-task networks is shown. Note that, grouped tasks in the branched architecture exhibit high RSA correlations, validating our searched configuration.

4.4 Task Pairing Case Study

Figure 3 shows a sample ResNet-50 encoder branching structure determined with our method. We compare the resulting task groupings with the Representational Similarity Analysis (RSA) matrix obtained from correlating encoded feature maps of the individual single-task networks, as proposed in [42]. The RSA suggests task groupings similar to ours, with the pairs SemSeg-PartSeg and Norm-Edge having high task affinities in both cases.

5 Conclusion

We presented BMTAS, an end-to-end approach for automatically finding efficient encoder branching structures for MTL. Individual task routings through a supergraph are determined by jointly learning architecture distribution parameters and neural operation parameters through backpropagation, using a novel resource-aware loss function. Combined, the routings form branching structures which exhibit high overall performance while being computationally efficient, as we demonstrate across several datasets (PASCAL-Context, NYUD-v2) and network backbones (ResNet-50, MobileNetV2). The proposed method is highly flexible and can serve as a basis for further exploration in MTL NAS.

Supplementary Material

A Training Settings

In this section, we describe the training setup used for experiments on PASCAL-Context. On NYUD-v2, the exact same setup was used, except that the number of training iterations was halved. Training and evaluation code for this project was written using the PyTorch library [42].

Data augmentation. We augment input images during training by random scaling with

values between 0.5 and 2.0 (in increments of 0.25), random cropping to input size (which was fixed to 512×512 for full-resolution PASCAL-Context) and random horizontal flipping. Image intensities are rescaled to the $[-1, 1]$ range.

Task losses. For semantic segmentation and human parts segmentation we use a cross-entropy loss (loss weights $\omega_t = 1$ and $\omega_t = 2$ in Equation 3, respectively), for saliency estimation a balanced cross-entropy loss ($\omega_t = 5$), for depth estimation a \mathcal{L}_1 loss ($\omega_t = 1$), for surface normal estimation a \mathcal{L}_1 loss with unit vector normalization ($\omega_t = 10$) and for edge detection a weighted cross-entropy loss ($\omega_t = 50$). For edge detection, the positive pixels are weighted with 0.95 and the negative pixels with 0.05 on PASCAL-Context, while on NYUD-v2 the weights are 0.8 and 0.2. ω_t for each task was found by conducting a logarithmic grid search over candidate values with single-task networks.

Optimization hyperparameters. Model weights θ are updated using Stochastic Gradient Descent (SGD) with momentum of 0.9 and weight decay 0.0001. The initial learning rate is set to 0.005 and decayed during training according to a ‘poly’ learning rate policy [4]. For the architecture distribution parameters α , we use an Adam optimizer [18] with learning rate 0.01 and weight decay 0.00005. We use a batchsize of 8 and 16 for ResNet-50 and MobileNetV2, respectively.

Architecture search. We update the supergraph sequentially for each task. Before the architecture search, we ‘warm up’ the supergraph by training each operation’s model weights θ (initialized with ImageNet weights) on the corresponding task only for 2000 iterations. The architecture distribution parameters α are initialized with zeros. During the search, we alternatively train α on 20% of the data and θ on the other 80%. This cycle is repeated until θ has received 40000 updates. Over the course of training, the Gumbel-Softmax temperature τ is annealed linearly from 5.0 to 0.1. Importantly, we use the batch-specific statistics for batch normalization during the α update phase and reset the batch statistics before training θ after every architecture change. Furthermore, to equalize the scale of candidate operations for the search, we disable learnable affine parameters in the last batch normalization of every operation. Finally, the momentum of the θ -optimizer is reset after every change to the architecture.

Branched network training. After the architecture search, the resulting branched network is retrained from scratch for 40000 iterations. The encoder network weights are initialized with ImageNet weights. For all operations that are shared between several tasks, we divide the learning rate by the number of tasks sharing, since those operations receive more updates.

B Implementation of Baselines

For FAFS [25], we pre-train a fully shared network on all the tasks and then calculate the task groupings in all layers greedily (starting from the last) according to the task affinity measure described in [25]. Note that for edge detection, we use the loss instead of the optimal dataset F-measure to determine sample difficulty. Finding branching structures via the BMN approach [17] involved training separate single-task networks and computing the Representational Similarity Analysis matrix from the resulting task-specific feature maps, exactly as described in the paper. Various branching structures can be found by exhaustively searching candidates among a reduced pool, containing all possible structures below a specified MAdds value. To keep the comparison fair, we trained the branched structures resulting from BMN and FAFS in exactly the same setting as ours.

Backbone	SemSeg \uparrow	PartSeg \uparrow	Sal \uparrow	Norm \downarrow	Edge \uparrow
MobileNetV2, [24]	62.10	54.88	66.30	14.88	69.50
MobileNetV2, ours	65.11	57.54	65.41	13.98	69.50
ResNet-50, [24]	68.30	60.70	65.40	14.61	72.70
ResNet-50, ours	70.43	63.93	67.34	13.39	74.10

Table S1: DeepLabv3+ performance in a single-task setting on PASCAL-Context using either MobileNetV2 or ResNet-50 as a backbone. We compare the performance obtained using our implementation with the results published in [24].

We implemented cross-stitch networks [24], NDDR-CNN [9] and MTAN [22] based on the code provided by the authors and information given in the papers. We use a similar training setup as the one described for our method, however we conducted a logarithmic grid search over learning rates for each baseline individually. For cross-stitch networks, applying one unit per feature tensor (as opposed to channel-wise) yielded more stable results. The weights of cross-stitch- and NDDR-CNN-units are initialized with $\alpha = 0.8$ and $\beta = \frac{0.2}{T-1}$, where T is the number of tasks. Both methods are applied on the fully pre-trained single-task networks. For MTAN with the MobileNetV2 backbone, we change the 3×3 convolutions in the attention modules to depthwise separable convolutions. In general, all ReLU activations are replaced with ReLU6 for MobileNetV2.

C Implementation Verification

To show that our implementations of DeepLabv3+ with the above mentioned backbones perform competitively for the tasks of interest, we compare in Table S1 our single-task performances on PASCAL-Context with published results in [24]. Note that a direct comparison is inconclusive even though the architectures are analogous, as the results in [24] are obtained with different training setups. Nevertheless, the numbers demonstrate that our single-task networks represent a strong baseline for comparison.

D Complementary Results

Due to space limitations, the performances of our method and simple baselines for a ResNet-50 backbone on PASCAL-Context (plotted on the right in Figure 2) are presented here in Table S2. For this setting, we choose not to report scores for cross-stitch networks [24], NDDR-CNN [9] and MTAN [22] since we were unable to obtain competitive performances for those approaches, despite the extensive learning rate grid-search.

References

- [1] Joachim Bingel and Anders Søgaard. Identifying beneficial task relations for multi-task learning in deep neural networks. *arXiv preprint arXiv:1702.08303*, 2017.
- [2] Felix JS Bragman, Ryutaro Tanno, Sebastien Ourselin, Daniel C Alexander, and Jorge Cardoso. Stochastic filter groups for multi-task cnns: Learning specialist and gen-

Model	MAdds ↓	SemSeg ↑	PartSeg ↑	Sal ↑	Norm ↓	Edge ↑	Δ_m [%] ↑
Single	345.7B	70.43	63.93	67.34	13.39	74.10	-
Shared	154.6B	68.24	62.18	65.16	14.98	71.90	-4.80
BMTAS-1	199.0B	68.17	62.36	65.64	14.09	72.20	-3.20
BMTAS-2	225.8B	66.92	62.93	65.82	13.70	72.90	-2.56
BMTAS-3	298.2B	69.58	64.36	66.68	13.65	73.00	-1.00

Table S2: Comparison of our method with simple baselines on PASCAL-Context using a ResNet-50 backbone. The resource loss weights λ for BMTAS were 2e-2, 5e-3 and 1e-3 respectively.

eralist convolution kernels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1385–1394, 2019.

- [3] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018.
- [4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
- [5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
- [6] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1978, 2014.
- [7] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. *arXiv preprint arXiv:1711.02257*, 2017.
- [8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [9] Yuan Gao, Jiayi Ma, Mingbo Zhao, Wei Liu, and Alan L Yuille. Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3205–3214, 2019.
- [10] Yuan Gao, Haoping Bai, Zequn Jie, Jiayi Ma, Kui Jia, and Wei Liu. Mtl-nas: Task-agnostic neural architecture search towards general-purpose multi-task learning. *arXiv preprint arXiv:2003.14058*, 2020.
- [11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

- [12] Ariel Gordon, Elad Eban, Ofir Nachum, Bo Chen, Hao Wu, Tien-Ju Yang, and Edward Choi. Morphnet: Fast & simple resource-constrained structure learning of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1586–1595, 2018.
- [13] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 270–287, 2018.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [15] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.
- [16] Menelaos Kanakis, David Bruggemann, Suman Saha, Stamatios Georgoulis, Anton Obukhov, and Luc Van Gool. Reparameterizing convolutions for incremental multi-task learning without task interference. *arXiv preprint arXiv:2007.12540*, 2020.
- [17] 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.
- [18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [19] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6129–6138, 2017.
- [20] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 82–92, 2019.
- [21] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.
- [22] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1871–1880, 2019.
- [23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [24] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and S Yu Philip. Learning multiple tasks with multilinear relationship networks. In *Advances in neural information processing systems*, pages 1594–1603, 2017.

- [25] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5334–5343, 2017.
- [26] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.
- [27] Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1860, 2019.
- [28] David R Martin, Charless C Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE transactions on pattern analysis and machine intelligence*, 26(5):530–549, 2004.
- [29] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016.
- [30] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016.
- [31] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Fast scene understanding for autonomous driving. *arXiv preprint arXiv:1708.02550*, 2017.
- [32] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [33] Hieu Pham, Melody Y Guan, Barret Zoph, Quoc V Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. *arXiv preprint arXiv:1802.03268*, 2018.
- [34] Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V Le, and Alexey Kurakin. Large-scale evolution of image classifiers. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2902–2911. JMLR. org, 2017.
- [35] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 4780–4789, 2019.
- [36] Clemens Rosenbaum, Tim Klinger, and Matthew Riemer. Routing networks: Adaptive selection of non-linear functions for multi-task learning. *arXiv preprint arXiv:1711.01239*, 2017.
- [37] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

- [38] Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. Latent multi-task architecture learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4822–4829, 2019.
- [39] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [40] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, pages 527–538, 2018.
- [41] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.
- [42] Simon Vandenhende, Stamatios Georgoulis, Bert De Brabandere, and Luc Van Gool. Branched multi-task networks: deciding what layers to share. *arXiv preprint arXiv:1904.02920*, 2019.
- [43] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. *arXiv preprint arXiv:2001.06902*, 2020.
- [44] Tom Veniat and Ludovic Denoyer. Learning time/memory-efficient deep architectures with budgeted super networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3492–3500, 2018.
- [45] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10734–10742, 2019.
- [46] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.
- [47] Sirui Xie, Hehui Zheng, Chunxiao Liu, and Liang Lin. Snas: stochastic neural architecture search. *arXiv preprint arXiv:1812.09926*, 2018.
- [48] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018.
- [49] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.
- [50] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.