MATTE: Multi-task multi-scale attention
Gjorgji Strezoski, Nanne van Noord, Marcel Worring
University of Amsterdam, Informatics Institute, Multix, Amsterdam, The Netherlands

A R T I C L E   I N F O
Communicated by Nikos Paragios

Keywords:
Multi-task learning
Matting
Visual Decathlon
Ubernet

A B S T R A C T
In this work, we propose a general method for learning task and scale based attention representations in Multi-Task Learning (MTL) for vision. It relies on learning and maintaining cross-task and cross-scale representations of visual information, whose interaction contributes to a symmetrical improvement across the entire task pool. Apart from learning data representations, we additionally optimize for the most beneficial interaction between tasks and their representations at different scales. Our method adds an attention modulated feature as residual information to the processing of each scale stage within the model, including the final layer of task outputs. We empirically show the effectiveness of our method through experiments with current multi-modal and multi-scale architectures on diverse MTL datasets. We evaluate MATTE on high and low level vision MTL problems, against MTL and single task learning (STL) counterparts. For all experiments we report solid performance improvements in both qualitative and quantitative performance.

1. Introduction

Behavioral science research (Szumowska et al., 2018) has shown that performing multiple related tasks simultaneously, often increases productiveness and throughput without harming performance. With similar reasoning, multi-task learning (MTL) is the ability of a computational model to optimize for multiple learning tasks in a single training cycle. MTL has been formally defined by Caruana (1997) as the approach to inductive transfer that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias (Thrun, 1996; Bakker and Heskes, 2003). As a training paradigm, MTL rests on the basis of sharing information between the tasks the model is training for, and using each individual control signal (loss) towards the benefit of the overall model performance. Sharing in MTL is performed using a parameter space, called the shared learning or shared parameter space which all of the tasks the model is training for tune using back-propagation (Caruana, 1997). This shared parameter space is what enables a single model to simultaneously represent and perform multiple tasks with different descriptive attributes.

MTL has numerous application domains. This ranges from simple recommender systems (Ning and Karypis, 2010; Zhang et al., 2019) which need to form a ranked list out of several meta-data modalities, to complex augmented reality (AR) systems (Lampropoulos et al., 2020) which need to perform efficient scene depth and surface normal estimations, instance segmentation, object recognition and alpha matting to render a semantically correct scene. Representing, or even capturing the information necessary for performing such tasks in the parameter space of a single, shared machine learning model, requires assembling information from multiple sensor types which can indicate luminescence, depth, material types, density, etc. The positive aspect of combining these information types is that it significantly narrows the gap between different modality representations. However, as a consequence the extent of possible MTL tasks grows rapidly (Zamir et al., 2018), inevitably leading to negative knowledge transfer (destructive interference) between tasks (Lee et al., 2016), performance averaging (Marquand et al., 2014) and non-converging models (Teichmann et al., 2018).

The problems (destructive interference, or performance averaging in an ever expanding parameter space) posed by the multitude of tasks, create the need for a moderated manner of training MTL models. In such models, shared information is passed through filters, such as attention mechanisms (Vaswani et al., 2017), before moving to its final output. Attention is a well known method of determining which part (spatial attention i.e. Woo et al. (2018)) of a feature map (channel attention i.e. Wang et al. (2017)) should be accentuated while learning/inference in order to enhance performance. In this paper, we propose MATTE, a method with which we transpose attention from the feature and spatial domain, to the task and scale domain. For example, when performing MTL on a task pool of semantic segmentation, depth, surface normal and trimap estimation, the depth and surface normal tasks benefit the most from sharing when their receptive fields are the same size, while trimap estimation can already benefit semantic segmentation features at a smaller scale by discarding most of the background surface from the course segmentation feature. In this way instead of maintaining feature correlations in isolation from the other
tasks and possible scales, we maintain the information across network layers and task specific branches until the final output. This operation fits well within the promise of MTL and definition of attention, as it enhances useful information sharing between tasks. In our attention formulation, the attention mechanism decides which channel, at which scale, gets how much resources in our shared learning space, before being subjected to the training control signal. By introducing MATTE to MTL, we preserve both the local and non-local affinity information at each scale. MATTE helps us to alleviate the effects of destructive interference, and to generate a more informative embedding before the final output layers take over.

To evaluate MATTE we conduct experiments on NYUD-v2 (Misra et al., 2016), SUN-RGBD (Song et al., 2015), KITTI (Uhrig et al., 2017), Visual Decathlon (Rebuffi et al., 2017), Ubernet (Kokkinos, 2017) and the alpha matting dataset (Rhemann et al., 2009). In the experiments conducted on these datasets we cover both high- and low-level vision tasks in an MTL model compared to STL and MTL approaches, and evaluated the impact of the influence between tasks. The experimental results show that any local and non-local affinities can be learned and propagated to form meaningful final representations of all the tasks across all available scales. Through the ablation study we also show that MATTE has a measurable influence on destructive interference and identify the distinct effect of every part of our approach. Across all experimental setups and datasets we report continuous performance improvements using MATTE in existing multi-task and multi-scale approaches.

We report our contributions in this paper in four aspects:

- We propose MATTE, an universal MTL attention method for combined cross-scale and cross-task learning that can be added to any multi-scale network architecture.
- With MATTE we measure and combat destructive interference between tasks, a known and still existing problem in MTL methods.
- We conduct extensive experiments over multiple datasets and benchmarks, reporting competitive or superior performance and qualitative findings.

2. Related work

Our experiments span the topics of semantic and instance segmentation, depth estimation, surface normal estimation, object detection and classification as individual tasks in an MTL flow orchestration. From an architectural and learning perspective, we draw inspiration from ideas in attention, gating and learning space partitioning. This section is structured to cover the most relevant works to these topics.

2.1. Multi-task learning in vision

General multi-task learning methods rest on sharing complementary information between the task specific branches of a shared model in order to regularize their outputs, boost performance and improve inference efficiency. The same principles apply to MTL in computer vision, as a vast body of research shows that sharing information through building a shared representation between tasks, benefits both high and low level vision problems (Guo et al., 2018; Zhang et al., 2018; Strezoski et al., 2019; Liu et al., 2019; Vandenhende et al., 2020a; Martin et al., 2004; Wang et al., 2016).

In low level vision problems, the models are dealing with dense pixel estimation. In such cases building a shared representation layer between tasks as the branching point of the model does not implicitly take into account the effect of sharing information between different scales within the feature space for one task. Vandenhende et al. (2020b) show that different receptive field sizes (and therefore different scales) in building such representations alter how tasks interact with each other. For example, a dilated kernel applied to a depth estimation map provides much more information about the scene semantics than an eroded one. As such, building a shared representation between a semantic segmentation model and depth estimation model would make more sense, and yield a larger benefit with a dilated receptive field. Wang et al. (2016) confirm a similar finding in their study on the combination of semantic segmentation and estimating surface normals. Given the benefits of distinctly sharing through the task or scale dimensions, in dense pixel estimation tasks we incorporate both through an attention mechanism and concatenated residual component.

High level vision problems such as image classification, object detection, image captioning or attribute prediction also benefit from MTL methods (Misra et al., 2016; Zhang et al., 2016; Liu et al., 2015; Li et al., 2014; Zhang et al., 2014b; Strezoski et al., 2019; Zamir et al., 2018; Ranjan et al., 2019; Teichmann et al., 2018; Jou and Chang, 2016; Alami Mejjati et al., 2018). The main distinction between the MTL methods applied in high and low level vision problems is the clear separation of the high level vision learning space into a task specific section and a shared one. The benefits of a partitioned learning space have been carefully studied (Mallya et al., 2018; Mallya and Lazebnik, 2018; Zhao et al., 2018; Zamir et al., 2018; Zhang et al., 2019a; Strezoski et al., 2019), especially in cases where unrelated tasks seem to mutually harm their learning process (Zhao et al., 2018). In this work, we preserve the partitioning of the learning space if already existing in architectures applied to high level vision problems, and encode it in the residual information forwarded to the processing of each scale.

2.2. Attention and learning space partitioning

As a manner of allocating the computational resources of the model towards the most informative signal (Olshausen et al., 1993; Itti and Koch, 2001, 2006; Larochelle and Hinton, 2010; Varwani et al., 2017), attention has been proven to help learning across many tasks in the multimedia, computer vision, and natural language processing domains. Regardless of the modality, attention mechanisms can appear
as spatial (Woo et al., 2018), cross-channel (Wang et al., 2017), cross-modality (Vandenheede et al., 2020b), cross-task (Liu et al., 2019), or combined as (Woo et al., 2018) show with a combination of spatial and channel attention. In our work, we demonstrate the effectiveness of spatial attention method between tasks, scales and ground-truth modalities.

In MTL, attention can appear as an implicit component of the training architecture (Strezoski et al., 2019; Yang and Hospedales, 2017; Li et al., 2014), or as a standalone component (Liu et al., 2019; Mallya et al., 2018; Zhao et al., 2018) trained alongside the tasks at hand. Some regularization properties in MTL networks, such as the random influence of one task over the learning process of others, can be considered analogous to Shake-Shake Regularization (Gastaldi, 2017) which adds a random variable to both the forward and backward flows during training. Similarly, partitioning in the shared parameter space can also be considered an attention mechanism as dedicated units correspond to task specific control signals, which limits or enhances the influence of affected channels (Yamada et al., 2018). As regularization between tasks is one of the core properties of MTL, in this paper we calculate attention and affinity maps between both tasks and scales across multi-task embeddings. The final embedding from MATTE is a result of the attention modulated joint contribution of all tasks across all scales, controlled at each scale step.

3. Method

Parameter sharing is the foundation on which MTL methods are built. How to share and distribute resources among task interaction is a central research point in the field. In this section, we introduce attention based sharing between tasks and scales, as well their combined auxiliary embedding information. Fig. 1 shows a high level overview of our method, and how the different task and scale information is combined with the help of auxiliary outputs at each scale step. The process of training and forwarding the auxiliary output information during training is displayed step by step in algorithm 1.

![Algorithm 1](image)

Algorithm 1 is repeated for each scale until the final output is reached. The final output gets the aggregated shared embeddings from each task at each prior scale when making the final predictions. Each scale has its own auxiliary output as in Fig. 2, from which a control signal is generated which trains our scale specific embeddings. In this way, we make sure that the shared task embedding is trained for each scale individually, and can measure the impact of the residual component propagated to the next scale.

In the sections below we use \( N \) as the set of tasks in each scale, \( S \) as the group of scales, and \( C, H, W \) as feature dimensions \((N \times S \times C \times H \times W)\). We denote \( \theta \) as the model parameters, \( F \) as the neural net, \( X \) as the input batch, \( L \) as the loss function and \( \delta \theta \) as the gradient. With this notation we can define \( f = F(X; \theta) \) as our model output. In a multi-task multi-scale environment the compound gradient is defined as \( \delta \theta = \sum_i \delta \theta_i + \sum_j \delta \theta_j \), where \( i \) is the task index, and \( j \) is the scale index. Defined as such, the parameter update rule is the standard compound gradient update defined as \( \delta \theta = \frac{\partial L}{\partial \theta} \), where \( \delta \theta \) is the cumulative gradient update for all tasks.

3.1. Cross-task attention

Our method works with multiple scales of the input data. The first step is to learn a shared representation for each task within the first scale. As all tasks train simultaneously without additional sampling strategies applied, the resulting feature will be marginally biased towards a prevalent task (the one experiencing the smallest loss). After up-sampling this feature to be propagated to the next scale, much of the weaker task information will be suppressed (Srivastava et al., 2014).

To remedy this, we attach a trainable attention block in the form of per task channel gating, which transforms our input from \( N \times S \times C \times H \times W \) to \( N \times S \times C_{\text{matte}} \times H \times W \), where \( C_{\text{matte}} \) represents the attention modulated features across all \( N \) tasks. The channel-wise transformation includes point-wise multiplication of the per task learned attention tensors for every channel. Different from other gating mechanisms (Strezoski et al., 2019; Woo et al., 2018; Mallya et al., 2018; Mallya and Lazebnik, 2018), our attention module preserves the task specific representation in the channel dimension similar to Liu et al. (2019). In fact, if attention is not applied, and the per task channels are only concatenated we empirically established that performance suffers despite the large resulting feature maps.

Following the attention block, a spatial pooling block (He et al., 2014) with a \( 1 \times 1 \) convolution output performs spatial pooling and dimensionality adjustment before propagating the embedding as a residual to the next scale or final output. Spatial pooling across the \( H \times W \) is suitable for cross-task operations as it does not interfere with the between task dynamics, nor does it affect the attention matrices.

It is important to note that in intermediate scale stages, the residual information propagates through the task specific heads before reaching the attention block. Every subsequent representation after the first
scale, carries the information from the previous scale with it. This residual information is also concatenated to the final modulated feature at the final scale as illustrated in Fig. 1. The diagram of our cross-task attention module in an individual scale is displayed in Fig. 3.

3.2. Cross-scale attention

Similarly to applying attention between the tasks specific representations at any individual scale, across scale attention can improve both efficiency and performance. The benefit is twofold as by discarding channels with redundant information the memory footprint of the overall model is reduced, and the resulting feature is distilled by retaining features persistent across scales. In this capacity, our cross-scale attention block has a channel gating and modulating role just before the final output layers of the task specific branches. This module converts the input from $N_i \times S \times C \times H \times W$ to $N_i \times S_{\text{main}} \times C_{\text{main}} \times H \times W$ dimensionality, where $C_{\text{main}}$ represents the attention modulated features across all $S$ scales per task.

This attention block is trained based on the interference introduced between combinations of the different scale representations of the input. A simple way to measure interference between scale representations is to follow the gradients produced from the outputs on task combinations and make sure they are not pulling in opposite directions. An example rule for applying attention to scales $S_1$ and $S_2$ of a single task which can positively or negatively influence each other:

$$A_{S_1, S_2} = \begin{cases} 
1 - |\Delta \theta_{S_1} - \Delta \theta_{S_2}|, & \text{sign}(\Delta \theta_{S_1} - \Delta \theta_{S_2}) > 0 \\
0, & \text{sign}(\Delta \theta_{S_1} - \Delta \theta_{S_2}) \leq 0 
\end{cases} \tag{1}
$$

Eq. (1) does not symmetrically apply to task level attention as in most cases the gradient scale between tasks can differ significantly. As final output, our cross-scale attention block outputs a feature map containing all the channels that satisfy the positive condition of Eq. (1). This final feature map is concatenated with the residual representations containing all the channels that satisfy the positive condition of Eq. (1). As depicted in Fig. 1, both the attention and task specific features are generated in a single forward pass and propagated to the tail of the model where the MATTE attention tensors modulate the resulting feature vector. With fanning out auxiliary outputs at the branching point of our models, we are able to apply MATTE to distinct signals (task, scale, and combined) and let the final feature selection be done just before the final output.

Having a multi-scale backbone increases the quantity of additional, sometimes redundant information which can harm the overall feature quality if not properly scaled. In the process of adding the residual information from any previous scale, we achieve a distillation effect by measuring the impact of the residual and allowing the control signal of the next scale to influence the transferred values.

3.3. Combined embeddings and residuals

Each task and each scale in our model, outputs separate and combined embeddings which are followed through to the final output layer. As depicted in Fig. 1, these embeddings can be considered horizontal across all tasks in a single scale, and vertical across all scales for a single task. Horizontally, for every task we measure the influence of the per task gradients with respect to the attention coefficients in the attention maps. This enables us to capture which tasks interact or respond well to the same data input. Vertically, for every task we can measure which embedding produces the best score and how much of the other scales content can be beneficial for the total performance metric. As depicted in Fig. 1, both the attention and task specific features are generated in a single forward pass and propagated to the tail of the model where the MATTE attention tensors modulate the resulting feature vector. With fanning out auxiliary outputs at the branching point of our models, we are able to apply MATTE to distinct signals (task, scale, and combined) and let the final feature selection be done just before the final output.

4. Experiments

In order to validate the contributions presented in this work, our experimental design takes into scope a number of classification, segmentation and dense prediction tasks. For this we consider a variety of datasets and benchmarks ranging from a multi-attributed based face dataset (CelebA), along a cross-dataset MTL benchmark (Visual Decathlon), to multi-output dense prediction (AlphaMatting, Adobe Image Matting, and VideoMatting.Com benchmark). To further explore the effects of the different attention components i.e. scale and task individually and combined, we perform an ablation study on the AlphaMatting dataset.

4.1. MTL in classification datasets

Multi-class classification datasets can be converted to a MTL by simply converting each existing class of the dataset into a binary task, assessing if a certain data sample belongs to a certain class. This experimental setup enables observing MTL in a homogeneous setting, where all data samples belong to the same dataset and have the same annotation form.

MNIST (Xiao et al., 2017) and CIFAR-10 (Krizhevsky et al., 2009) constitute the proof of concept part of our experimental design as they are well established benchmarks and provide simple indication of how different hyper-parameter setups tend to affect the method. For both dataset we define ten binary classification tasks and evaluate the accuracy, precision, and recall scores.

UCSD Birds (Wah et al., 2011) is a dataset that provides 11,788 bird images over 200 bird species with 312 binary attribute annotations. For state of the art comparison, we compare on ten target attributes obtained with spectral clustering using the FSIC as the similarity measure (Alami Mejjati et al., 2018). As we incrementally increase the selected number of attributes we define sets of 50, 100, 200, and 312 binary classification tasks for each of them. For this dataset the training and testing set have equal sizes and distributions. The attributes are sampled according to the ten attribute selection in Alami Mejjati et al. (2018) for the 10 task experiment and in order of the original annotation file for the rest of the experiments.

CelebA (Liu et al.) consists of more than 200,000 face images with binary annotations on 40 face attributes related to age, expression, decoration, etc. The first 10 attributes from Zhao et al. (2018) are selected for the 10 task experiment as more related to face appearance. We additionally report the results on 40 attributes to compare MATTE to Zhao et al. (2018) in a classification setting.

UT-Zappos50K (Yu and Grauman, 2017) is a large shoe dataset consisting of more than 50,000 catalog images collected from the web. This dataset contains four attributes of interest for our experiment, namely shoe type, suggested gender, height of the heel, and the shoe closing mechanism. As defined in Zhao et al. (2018), we define 4 classification tasks for a small scale test of MATTE over a real world dataset using the identical train, validation, and test splits from Veit et al. (2017), Zhao et al. (2018).

4.2. MTL in Matting datasets

Alpha matting as a computer vision task consists of three distinct processes (background/foreground segmentation, trimap estimation and image matting) resulting in a final output. The challenging aspect to tackle is the co-dependent nature of the task outputs leading to the final output — the alpha matte mask. This dependency between tasks can bring to light aspects of destructive interference in MTL learning and also show the importance of scale and receptive fields in MTL problems. For our MTL setup we define three tasks, semantic segmentation, trimap estimation and matting across three datasets.

The AlphaMatting Benchmark is a popular image matting benchmark, with eight objects for which trimaps and alphas are available.
for training. In addition, there is a well defined evaluation protocol with a controlled set of testing images. Image matting as a task has a trimap image requirement, which is a pixel mapping of background, foreground and area of uncertainty pixels which we provide as an auxiliary input from the output of our trimap estimation task. The evaluation metrics for these tasks are explained in Section 4.4.

The Adobe Image Matting Benchmark The Adobe Image Matting benchmark offers a well structured dataset based on Pascal VOC with trimaps, alpha mattes and segmentation masks. In the same manner as with the AlphaMatting benchmark, our MTL construction is based on three tasks: alpha matting, background/foreground segmentation and trimap estimation. The evaluation metrics for these tasks are explained in Section 4.4.

The VideoMatting project (Erofeev et al., 2015) is the first public video matting benchmark which offers public training sequences and per frame ground-truth, as well as a public evaluation engine. For this dataset we process the videos in their native frame-rate and output a per frame ground-truth, as well as a public evaluation engine. For this purpose, the method could bridge the missing annotation gap.

Table 1

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MNIST</th>
<th>CIFAR-10</th>
<th>Zappos50K</th>
<th>CelebA</th>
<th>CelebA (Full)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Tasks</td>
<td>10</td>
<td>10</td>
<td>4</td>
<td>10</td>
<td>40</td>
</tr>
<tr>
<td>Approach</td>
<td>Accuracy</td>
<td>Precision</td>
<td>Recall</td>
<td>Accuracy</td>
<td>Precision</td>
</tr>
<tr>
<td>Cross-stitch</td>
<td>96.1±14</td>
<td>91.4±1.02</td>
<td>86.1±0.23</td>
<td>98.5±0.17</td>
<td>91.6±1.18</td>
</tr>
<tr>
<td>Modulation</td>
<td>98.0±2.04</td>
<td>91.0±1.14</td>
<td>86.1±0.54</td>
<td>82.2±1.13</td>
<td>93.7±1.04</td>
</tr>
<tr>
<td>Res2Net</td>
<td>95.1±13</td>
<td>109.9±0.62</td>
<td>78.9±0.22</td>
<td>72.2±1.33</td>
<td>1.13</td>
</tr>
<tr>
<td>MATTE (Res2Net)</td>
<td>98.6±0.26</td>
<td>93.2±0.11</td>
<td>85.7±0.03</td>
<td>96.0±0.03</td>
<td>92.1±0.08</td>
</tr>
</tbody>
</table>

These are metrics defined to capture the temporal dependencies across the processed frames. Their correlations with respect to human perception and individual frame matting metrics is explained in detail in Rheinm (2009). The image matting tasks are evaluated using the SSD and MSE metrics.

We evaluated our attention approach in multiple classification settings in a classic MTL setup. The goal of our evaluation is to determine the effect of learning a hybrid form of attention across tasks and scales has on the final outcome in a MTL setting. In addition, we detect and correlate inherent connections between the tasks when the model is asked to combine channels across the feature maps extracted at different scales.

4.5. Experimental setup

For our experimental setup we use a MS-COCO pre-trained HR-Net (Sun et al., 2019) and Res2Net (Gao et al., 2019) as our multi-scale backbone implementations in Pytorch 1.8 and Cuda 10.2. Our models were trained on 4 Nvidia 1080ti GPUs with LaPack version 3.9.0.1 The complexity of MATTE is approximately the same as the native complexity of the baseline models it is applied to. For example in a 50 layer network (ResNet-50) whose number of parameters is around 25M and the number of FLOPs for an image of 224 × 224 pixels is 4.2G, the multi-scale equivalent with MATTE has 500K additional parameters and the same FLOP count. For training a multi scale backbone in MTL mode, we adopted the low memory back-propagation method from Chen et al. (2016), Kokkinos (2017) and implemented an exception handling class for empty annotation values which persist the weight values when annotations are missing per class. Our exception handling enables to skip the nullable calculations which results in a 5% decrease in training time over conventional MTL training.

As in most MTL training pipelines (Guo et al., 2018; Zhang et al., 2018; Strezoski et al., 2019; Liu et al., 2019; Vandenheede et al., 2020a; Martin et al., 2004; Wang et al., 2016), our experiments seem to benefit from lower batch sizes. For each experiment, we use the maximum batch size to the memory capacity of our system. With respect to learning rates, the best performance and fastest convergence rates we experienced with smaller initial learning rates set to 5e-4, similar to Kokkinos (2017).

4.6. Ablation study

Our proposed attention mechanism depends on learning the combination of tasks and scales with the best mutual affinity. To establish a quantifiable metric of the influence of the chosen scale channels against the task channels, we conduct a qualitative and quantitative

1 https://github.com/leoxiaobin/deep-high-resolution-net.pytorch
study on the AlphaMatting dataset with three tasks with and without the individual attention modules. Our first experiment is founded on a bare multi-scale MTL backbone with a branch per output. In our second experiment we rely on the scale affinities as the main attention component, and in the third experiment rests on the task combinations. Finally we compare the three ablation approaches against our main experiment containing the same backbone and the combination of both scale and task attention factor combinations. Our ablation study continues in the UberNet experimental setup where the subtraction of selected tasks from the MTL pipeline enables us to observe the effect of the between task influence with or without MATTE attention, further described in Section 5.4.

5. Results

As defined in Section 4.1, experiment one, we start our evaluation with a sanity check experiments on a small proof of concept dataset (MNIST). The intuitive tuples of tasks such as 1 and 7, 6 and 9, 2 and 5 or 3 and 8 quickly form after completing only 2 epochs. This behavior of the network confirms that the grouping of tasks makes visual sense and can be attributed to visible similarities in the input images (see Fig. 4).

5.1. Evaluation on UberNet

Our UberNet setup evaluation creates the opportunity of observing MATTE in a diverse annotation setting where labels can be missing in various tasks during training. The overall performance in Table 7 shows a consistent improvement along almost all tasks when attention is applied in both single and multi scale settings. In Table 7 we report an improvement in overall performance as well as in individual tasks singled out (Tables 2–4). It is through this experimental setup that we identify the influence MATTE has on destructive interference as well. As we report our performance, we are dealing with an increased feature

---

**Table 2**

Mean Average Precision on the PASCAL VOC 2007 test set in STL, 2 task (Object detection and Semantic Segmentation) and 7 task MTL scenarios. We report results with the Res2Net backbone with and without MATTE.

<table>
<thead>
<tr>
<th>Method</th>
<th># of tasks</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-RCNN MS COCO + VOC 2007</td>
<td>1</td>
<td>78.8</td>
</tr>
<tr>
<td>UberNet STL</td>
<td>1</td>
<td>78.7</td>
</tr>
<tr>
<td>UberNet MTL (Sem. Seg + Det.)</td>
<td>2</td>
<td>80.1</td>
</tr>
<tr>
<td>UberNet MTL</td>
<td>7</td>
<td>77.8</td>
</tr>
<tr>
<td>TAG (Single Scale)</td>
<td>7</td>
<td>84.5</td>
</tr>
<tr>
<td>Res2Net (Single Scale)</td>
<td>1</td>
<td>78.7</td>
</tr>
<tr>
<td>MATTE Res2Net (Single Scale)</td>
<td>1</td>
<td>81.2</td>
</tr>
<tr>
<td>Res2Net (Single Scale) (Sem. Seg + Det.)</td>
<td>2</td>
<td>83.5</td>
</tr>
<tr>
<td>Res2Net (Multi Scale) (Sem. Seg + Det.)</td>
<td>2</td>
<td>84.1</td>
</tr>
<tr>
<td>Res2Net (Multi Scale)</td>
<td>7</td>
<td>82.8</td>
</tr>
<tr>
<td>MATTE (Multi Scale) (CT)</td>
<td>7</td>
<td>82.5</td>
</tr>
<tr>
<td>MATTE (Multi Scale)</td>
<td>7</td>
<td>84.9</td>
</tr>
</tbody>
</table>

**Table 3**

Results on the Surface Normal Estimation task from the UberNet setup. On this dataset (NYU-v2) we report the Mean and Median Angle distance (in radians) and the percentage of valid pixels existing in ~11.25, 22.5 and 30 degrees from the ground-truth. The final four rows correspond to a plain multi scale Res2Net, MATTE with Cross-Task (CT) Attention, MATTE with Cross-Scale Attention (CS) and MATTE with both.

<table>
<thead>
<tr>
<th>Method</th>
<th>Tasks</th>
<th>Mean</th>
<th>Median</th>
<th>11.25</th>
<th>22.5</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG-Cascade</td>
<td>1</td>
<td>22.2</td>
<td>15.3</td>
<td>38.6</td>
<td>64.0</td>
<td>73.9</td>
</tr>
<tr>
<td>VGG-Fused</td>
<td>1</td>
<td>27.9</td>
<td>16.6</td>
<td>37.4</td>
<td>59.2</td>
<td>67.1</td>
</tr>
<tr>
<td>UberNet STL</td>
<td>1</td>
<td>23.9</td>
<td>18.1</td>
<td>29.8</td>
<td>59.7</td>
<td>71.9</td>
</tr>
<tr>
<td>UberNet MTL</td>
<td>7</td>
<td>26.7</td>
<td>22.0</td>
<td>24.2</td>
<td>52.0</td>
<td>65.9</td>
</tr>
<tr>
<td>TAG (Single Scale)</td>
<td>1</td>
<td>23.2</td>
<td>17.8</td>
<td>29.2</td>
<td>53.8</td>
<td>65.3</td>
</tr>
<tr>
<td>MATTE (Single Scale)</td>
<td>1</td>
<td>22.8</td>
<td>16.5</td>
<td>30.9</td>
<td>55.0</td>
<td>73.2</td>
</tr>
<tr>
<td>MATTE (Multi Scale)</td>
<td>7</td>
<td>20.7</td>
<td>16.6</td>
<td>39.2</td>
<td>63.9</td>
<td>75.0</td>
</tr>
<tr>
<td>TAG (Single Scale)</td>
<td>7</td>
<td>20.0</td>
<td>14.8</td>
<td>39.3</td>
<td>64.0</td>
<td><strong>77.0</strong></td>
</tr>
</tbody>
</table>

Note that a task cannot pair with itself.
entropy in experimental setups such as UberNet. In such cases any model will clearly benefit from having more free parameters, additional nonlinearities and even larger batch sizes or strictly controlled learning rates (Chen et al., 2018). This means that training a larger model, or expanding the ones proposed in this work is likely to result in an improved overall performance. In Fig. 5 we visually showcase the combined output for all tasks in the UberNet setup.

### 5.2. Evaluation on matting

Table 1 shows results for both heterogeneous and homogeneous MTL setups. Adding MATTE before computing the aggregated feature map for the output benefits both classification performance and the learned attention matrices for the tasks at hand. The measured performance with and without our attention modules shows the learned attention modulation does not filter through the scale specific representations and only represents the task to task affinities. The tasks themselves carry either a positive or negative signal for all of their counterparts, which can be learned in a differentiable manner. This signal is positively amplified if combined with the distinct per scale representations. The key finding of this experiment with classification tasks is that attention maps can be learned with a combination of scale representations. The key finding of this experiment with classification tasks is that attention maps can be learned with a combination of scale representations. The key finding of this experiment with classification tasks is that attention maps can be learned with a combination of scale representations. The key finding of this experiment with classification tasks is that attention maps can be learned with a combination of scale representations. The key finding of this experiment with classification tasks is that attention maps can be learned with a combination of scale representations. The key finding of this experiment with classification tasks is that attention maps can be learned with a combination of scale representations.

### 5.3. Evaluation on the VD challenge

The VD challenge poses a different heterogeneous type of MTL problem, where each dataset is defined as a task. We implemented two scale based MTL approaches (Strezoski et al., 2019; Zhao et al., 2018) to fit a multi-scale backbone MTL flow with Res2Net for a fair comparison. Compared to the rigid constraint of channel gating in Strezoski et al. (2019), our hybrid attention form is more flexible when it comes to influences from multiple scales in the final embedding and produces superior overall results. The latest submission on the official VD evaluation engine has the single scale (Zhao et al., 2018), which improves significantly with transitioning to a multi-scale backbone, but plateaus quickly after saturating the learnable modulation adjustment parameters. Table 6 shows the individual and overall scores, where MATTE scores the best overall challenge metric. For more qualitative findings please refer to the appendix.

### 5.4. Destructive interference from surface normal estimation

When a pair of tasks react in an opposite manner to the same input, most often it is because the control signal is providing competing or even contradicting gradient directions over a shared layer of units. This phenomenon, known in Physics as destructive interference, occurs when two waves of equal frequency and opposite phases occupy the same physical space and cancel each other out. Similarly, in MTL the calculated gradients of two tasks with opposite directions destroy each other’s influence over the trainable parameter space.

Destructive interference is clearly observable in our experimental setup when introducing the surface normal estimation problem to the pool of tasks the network performs. Table 8 shows an overall degradation in performance once the Surface Normal Estimation task is added in rows 3, 4, 5 and 7. It also notes a spike in performance in row 6 once learned in isolation (STL). Our reasoning relates to the performance

<table>
<thead>
<tr>
<th>Method</th>
<th># of tasks</th>
<th>ODS ↑</th>
<th>OIS ↑</th>
<th>AP ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiScale sPh</td>
<td>1</td>
<td>0.726</td>
<td>0.757</td>
<td>0.696</td>
</tr>
<tr>
<td>HED - Fusion</td>
<td>1</td>
<td>0.790</td>
<td>0.808</td>
<td>0.811</td>
</tr>
<tr>
<td>UberNet MTL</td>
<td>7</td>
<td>0.785</td>
<td>0.805</td>
<td>0.837</td>
</tr>
<tr>
<td>UberNet STL (sPh setup)</td>
<td>1</td>
<td>0.815</td>
<td>0.835</td>
<td>0.862</td>
</tr>
<tr>
<td>TAG (Single Scale)</td>
<td>7</td>
<td>0.830</td>
<td>0.838</td>
<td>0.874</td>
</tr>
<tr>
<td>Res2Net (Single Scale)</td>
<td>7</td>
<td>0.789</td>
<td>0.821</td>
<td>0.866</td>
</tr>
<tr>
<td>MATTE (Multi Scale) (CT)</td>
<td>7</td>
<td>0.801</td>
<td>0.827</td>
<td>0.868</td>
</tr>
<tr>
<td>MATTE (Multi Scale) (CS)</td>
<td>7</td>
<td>0.819</td>
<td>0.834</td>
<td>0.875</td>
</tr>
<tr>
<td>MATTE (Multi Scale)</td>
<td>7</td>
<td>0.836</td>
<td>0.840</td>
<td>0.886</td>
</tr>
</tbody>
</table>
issues reported in Kokkinos (2017) as well. The surface normal estimation problem of the UberNet setup poses a harder challenge in an MTL setup as the label definition of the input is of a more continuous and geometric nature. The result of that ground-truth difference is that the backbone networks with their ImageNet and Coco pre-trained weights do not pose a relevant feature bed without additional nonlinearities and heavy fine-tuning for surface normals.

Using MATTE clearly allows for competitive performance in cases when destructive interference occurs, because the attention modules themselves do not pose a relevant feature bed without additional nonlinearities and heavy fine-tuning for surface normals.
task improvement. Table 8 shows that the degradation in performance of the surface normals task done in STL mode, compared to the native task performance done in an MTL setting is 2.3% in row 10 compared to row 6, but significantly smaller — 4.7% (between row 6 and 7) than an MTL setup without MATTE and an explicit task prioritization approach (Guo et al., 2018).

6. Conclusion

In this work we show the power of a combined scale and task-wise attention in multi-scale MTL models. For our primary contribution, we propose a MTL attention mechanism, MATTE, that leverages the combined representational power of multi-scale backbones and MTL models for learning superior representations and boosting performance. We achieve such behavior by increasing the influence of channels across tasks and scales beneficial to performance, using a trainable non-linearity over task and scale affinity tensors. By applying attention in such manner MATTE combats destructive interference across tasks and scales in multi-scale MTL. Our qualitative and quantitative findings suggest a clear benefit of using MATTE as an attention mechanism in MTL in both high level tasks such as image classification, and dense output tasks such as segmentation, trimap generation and matting in both videos and images.

Our dense output experiments with segmentation and matting tasks give a more intuitive outlook on the effect of MATTE as an attention mechanism, and also treat trimap estimation as a learnable task rather than a manual transformation. Having this type of dynamic trimap estimation, leads us closer to a true end to end matting approach, where the size of the uncertainty area in a trimap is dictated by the matting model. While having multiple trimap candidates and iteratively applying the trimap to the refinement of the final matting result works very well, it still generates redundant data and model iterations to produce the final result. With MATTE the conventional ensemble of trimap approach to matting can be replaced by a trimap generated to produce best matting results with the underlying model, reducing the redundant trimap generation, reducing the number of I/O operations if interim results are stored and producing competitive or better performance.

Our experiments with the Ubernet setup helped us observe how our method would cope with missing annotations in a MTL dataset, but also measure the effect an attention mechanism such as MATTE has on destructive interference. Instead of freezing training for a branch with missing annotations, it can be inherently handled by the attention coefficients masking the resulting gradient for the specific branch. Attention is also a promising way of addressing destructive interference by making the model aware of how tasks interact with each other. In MTL setups with multi scale feature extraction, instead of amplifying the negative effects, using MATTE on the multi scale features can help in learning a superior representation.

The true strength of MTL over STL approaches is best visible when there is a structured sharing approach, in a tightly related set of tasks over a descriptive dataset. We conclude that even in tasks as related as segmentation, trimap estimation and matting negative interference can affect the final outcome if the shared representation is not properly modulated. With MATTE, we present an approach that can enables multi-scale MTL approaches to achieve the promise MTL models are built for.

CRediT authorship contribution statement

Gjorgji Strezoski: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft. Nanne van Noord: Validation, Writing – review & editing, Supervision. Marcel Worrin: Validation, Resources, Writing – review & editing, Supervision, Project administration, Funding application.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References


