Cross-Domain Multi-task Learning for Object Detection and Saliency Estimation

Apoorv Khattar
TCS Research
a.khattar@tcs.com

Srinidhi Hegde
TCS Research
sri.hegde@tcs.com

Ramya Hebbalaguppe
TCS Research
ramya.hebbalaguppe@tcs.com

Abstract

Multi-task learning (MTL) is a learning paradigm that aims at joint optimization of multiple tasks using a single neural network for better performance and generalization. In practice, MTL rests on the inherent assumption of availability of common datasets with ground truth labels for each of the downstream tasks. However, collecting such a common annotated dataset is laborious for complex computer vision tasks such as the saliency estimation which would require the eye fixation points as the ground truth data. To this end, we propose a novel MTL framework in the absence of common annotated dataset for joint estimation of important downstream tasks in computer vision - object detection and saliency estimation. Unlike many state-of-the-art methods, that rely on common annotated datasets for training, we consider the annotations from different datasets for jointly training different tasks, calling this setting as cross-domain MTL. We adapt MUTAN [3] framework to fuse features from different datasets to learn domain invariant features capturing the relatedness of different tasks. We demonstrate the improvement in the performance and generalizability of our MTL architecture. We also show that the proposed MTL network offers a 13% reduction in memory footprint due to parameter sharing between the related tasks.

1. Introduction

Deep learning has made significant strides in enabling intelligent systems to learn complex tasks. However, the absence of a suitable dataset with all the task specific annotations makes it harder to learn a wide range of tasks simultaneously. Thus, this poses a serious challenge in multi-task learning (MTL) [7] paradigm which necessitates the datasets to have ground truth annotations for all the tasks in the task set. Therefore, we enable the MTL frameworks to utilize the ground truth annotations from different domains for learning different tasks and inspired by [29], we call this setup as the *cross-domain* multi-task learning.

In this work, we propose an MTL framework for the joint optimization of object detection and visual saliency prediction tasks. The rationale behind such a task set stems from the direct applications in salient object detection [48], situated visualization [22], high dynamic range imaging [35, 27], and remote sensing [13]. Furthermore, in an MTL setting to improve the network performance, it is important for the tasks in the task set to facilitate each other. This aspect is conveyed by the task relatedness [2]. Thus, we analyze the relatedness between different complex computer vision tasks using the Representation Similarty Analysis (RSA) [15]. Figure 1 shows the relatedness scores between the different tasks. Our task set consists of Object Detection and Saliency Estimation which has a very low relatedness score and, thus, is challenging to jointly optimize. We also demonstrate that the relatedness score dips even further in a cross-domain setting (See section 5.5).

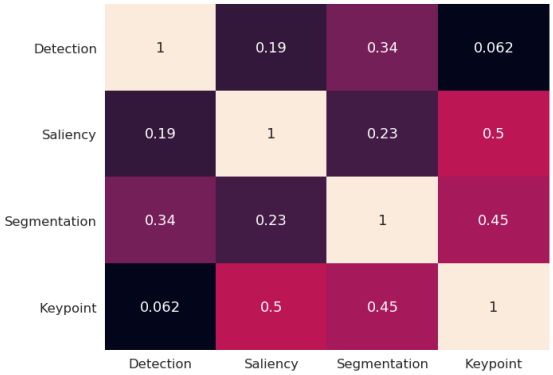

Figure 1. Task similarity matrix. We compute the task relatedness among different computer vision tasks such as object detection (Detection), semantic segmentation (Segmentation), saliency estimation (Saliency), and human pose keypoint detection (Keypoint). We use MS-COCO [33] dataset and models SSD300 [36] for Detection, MSI-Net [30] for Saliency and Segmentation, and Xiao et al. [49] for Keypoint. All models use ResNet-50 [21] as feature extractor. We observe that the pairs (Saliency, Detection) and (Detection, Keypoint) are least related and hence are challenging to learn in an MTL setting.

Estimating the image patches which a human eye finds salient is challenging for a DNN to learn, given the subjective nature of the task. Therefore, ground truth annotations

for saliency are hard to collect, in comparison to other visual annotations. Visual saliency of a given image is represented using a saliency map. The intensity assigned to each pixel in the saliency map determines the visual salience of the respective corresponding pixel of the image in the scene. Thus, our work is useful in joint learning of complex tasks with special data requirements.

Despite the state-of-the-art performance of Deep Neural Networks (DNNs) in downstream tasks in vision, NLP, and speech, DNNs are often overparameterised. In the light of applications demanding real-time performance, the efficiency and compactness play an important role. Therefore, we analyze the effect of parameter sharing in MTL architectures, for reducing the model parameters and the memory footprint, in contrast to the standalone DNN architectures. To summarize, our key contributions in this work are:

- We propose an MTL architecture that jointly performs saliency estimation, and object detection where the data originates from different domains.
- We adapt a bilinear kernel function for learning domain invariant features by learning the relation between different feature vectors of images sampled from different domains.
- We perform extensive evaluation of the proposed network in a cross-domain setup on accuracy, performance, memory footprint, and also analyse the effects of cross-domain setup on the RSA based task relatedness.

2. Related Work

2.1. Object Detection

Object detectors are broadly categorized into region-based (2-stage) and unified (1-stage) object detectors. Region-based object detectors such as [18, 12, 32] compute candidate regions, or the region proposals, with the possible presence of object of interest and then determine the class labels for the computed region proposals. In the unified object detectors such as [40, 36, 16] are the end-to-end architectures that directly predict the class probabilities and bounding box offsets of the objects found in an entire input image. Since we are interested in the compute- and the memory-efficient object detection, we focus on the unified frameworks as, unlike region-based frameworks, they circumvent region-proposal computation reducing computational complexity. YOLOv2 [40] and SSD [36] are the widely used object detectors. YOLOv2 treats the object detection as a regression problem and SSD integrates the region-proposal concept into a unified framework using a multi-scale CNN architecture. Fu et al. [16] extend the SSD architecture to enhance performance and accuracy. Recently, Carion et al. [6] employ a transformer encoder-decoder architecture [46] by posing object detection as set-prediction problem. This method provides significant memory footprint reduction retaining competitive accuracy.

2.2. Saliency Prediction

Traditionally, visual saliency prediction models were inspired by biological features that captured low level features in image such as color, edges, texture and semantic abstractions of certain objects of interests [1, 20]. Recent advancement in deep learning techniques and large annotated datasets popularized data-driven approaches. Cornia et al. [11] propose a novel architecture using attentive convolutional LSTM for saliency estimation and also introduce a prior module to account for the center bias in human eye fixations. Zhang et al. [52] propose a multi-path recurrent network which uses spatial attention mechanisms for accurate separation of the salient objects from the background. Few of the works [37, 8] have shown the effectiveness of generative adversarial networks (GAN) [19] to achieve competitive results on MIT Saliency Benchmark [5].

2.3. Multitask Learning

Multitask Learning (MTL) is a learning paradigm involving joint optimization of multiple task objective using shared knowledge. Such a joint optimization is shown to boost model performance through the introduction of inductive bias [7]. MTL is extensively used in many computer vision applications such as scene understanding [14, 50, 28], age and gender classification [17], and many more [51]. Broadly, MTL techniques can be divided into two categories. The first category focuses on the different ways of parameter sharing between different tasks. Some of the works in this category [14, 28, 10] focus on rich task agnostic feature extractors, known as encoders, which are used by task specific streams to jointly minimize the weighted loss for all tasks. Conversely, few works [50, 53, 45] also propose models that refine their initial task-specific predictions using shared neural network, known as decoder. The second category of works [44, 10, 28] focus on balancing the joint learning of tasks to avoid dominance of one or more tasks while training the network. Unlike our work, all of the aforementioned methods assume a common dataset with annotations available for each of the tasks. In our work, the proposed MTL architecture is inspired by Blitznet [14] that jointly performs object detection and semantic segmentation.

2.4. Cross-Domain Learning

Cross-Domain Learning focuses on learning a particular task from inputs sampled from multiple probability distributions, which is formally defined in [29]. Ren et

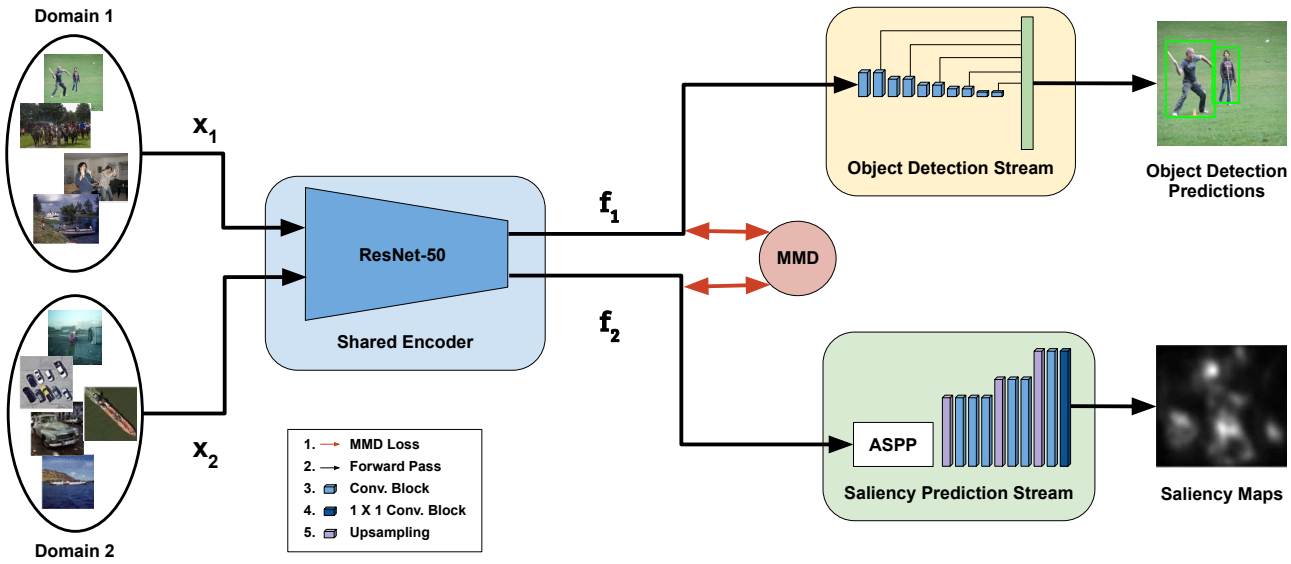

Figure 2. Architectural overview of the proposed MTL network in a cross domain setting. We pass the input images (x_1, x_2) from different domains (or datasets) through a backbone CNN to extract high-level image features (f_1, f_2). Multiple task specific streams take these features from the shared backbone CNN and predicts the outputs (the saliency map and the detected objects' classes and bounding boxes) for respective tasks. We use an MMD loss between f_1 and f_2 for learning domain agnostic features. In the object detection stream, we use multi-scale feature maps similar to SSD [36]. For saliency estimation, we use a Atrous Spatial Pyramid Pooling module (ASPP) [9], to get high resolution dense prediction, followed by a CNN decoder network.

al. [41] explore cross-domain learning in MTL setup to jointly learn depth, surface normal, and object contours using both real and synthetic images. Recently, [34] follow a cross-domain MTL approach to develop view-invariant and modality-invariant models for jointly learning activity specific human action recognition. In our work, we extend the idea of cross-domain MTL to solve object detection and saliency estimation tasks and analyse the role of task relatedness in this setup.

3. Proposed Work

We propose a novel architecture for joint object detection and saliency estimation that aids in better scene understanding. Figure 2 shows an overview of the proposed network architecture which uses a shared backbone convolutional neural network (CNN). Furthermore, the annotated data for the two tasks belongs to different domains. Thus, it becomes essential for backbone CNN to learn domain invariant features. In the following subsections, we discuss the details of cross-domain MTL and module-wise objective functions.

3.1. Cross-Domain Multi-Task Learning

We now look at the cross-domain MTL problem formally. Consider $D_S = \{x_1 ... x_n\}$ and $D_T = \{y_1 ... y_m\}$ as the source and the target distributions respectively. Our goal is to learn a feature extractor, ϕ, that transforms the input image space, $\mathcal{X}$, to a Reproducible Kernel Hilbert Space (RKHS), $\mathcal{H}$, such that the distance between D_S and D_T is minimized. Inspired by its wide usage in domain adaptation, we measure the distance between D_S and D_T with Maximum Mean Discrepancy (MMD) [4] defined as,

$$MMD(D_S, D_T) = \left\| \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) - \frac{1}{m} \sum_{i=1}^{m} \phi(y_i) \right\|_{\mathcal{H}} \quad (1)$$

As explained in [4], RKHS assumption on the feature space aids in balancing the tradeoff between *overfitting* and *underfitting* through the MMD statistic. For convenience, we minimize $MMD(D_S, D_T)^2$ instead of $MMD(D_S, D_T)$. Thus, applying a suitable kernel function k, we have,

$$\begin{aligned} MMD(D_S, D_T)^2 &= \left\| \mathbf{E}_{D_S}[\phi(x)] - \mathbf{E}_{D_T}[\phi(y)] \right\|_{\mathcal{H}}^2 \\ &= \mathbf{E}_{D_S, D_S}[k(x_i, x_j)] + \mathbf{E}_{D_T, D_T}[k(y_i, y_j)] \\ &\quad - 2\, \mathbf{E}_{D_S, D_T}[k(x_i, y_j)]. \end{aligned} \quad (2)$$

Generally, for domain adaptation, an RBF kernel is a popular kernel choice. However, inspired by their application in visual question answering (VQA), we employ a kernel function based on the bilinear models [43]. VQA

requires a deep understanding of relationships between the image and the text features. To this end, [3] propose MUTAN, a novel technique to fuse the image and the text features by computing a weighted pairwise product between the elements of the two features.

Similarly in domain adaptation, to learn domain invariant features the kernel function must capture the relationship between the source and target features. In an RBF kernel, the higher order represent these combinations however since their magnitude is less these relations are not properly captured. Hence, inspired by MUTAN [3], we use a learnable bilinear kernel function $k : \mathcal{R}^n \times \mathcal{R}^n \rightarrow \mathcal{R}$, in Eq. 2, given by,

$$k(z, z') = z^T A z' + c, \tag{3}$$

where $A = W^T W + I$, W is a learnable non-zero weight matrix, I is the identity matrix and c ($c >= 0$) is a learnable scalar. Thus, we can see that k is a valid kernel function which preserves RKHS assumption as A is positive definite. We, now, define the loss term for constrained MMD as,

$$\mathcal{L}_{MMD} = MMD(D_S, D_T)^2 - log(\|W\|_2) - log(c) \tag{4}$$

Note that we use the extra regularization terms for avoiding the W to saturate to zero and c to remain non-negative.

3.2. Multi-Task Objective Function

The object detection stream predicts the bounding box offsets and class scores for each object in the image. The task-specific loss for this stream is given by,

$$\mathcal{L}_{detect}(p_c, g_c, p_b, g_b) = \mathcal{L}_{class}(p_c, g_c) + \mathcal{L}_{loc}(p_b, g_b), \tag{5}$$

where $\mathcal{L}_{class}$ is the cross entropy loss with p_c and g_c as the predicted and ground truth class respectively. $\mathcal{L}_{loc}$ is the localization loss that is the L1 loss between the predicted and ground truth bounding box offsets, p_b and g_b.

For saliency estimation, we use the following loss function,

$$\mathcal{L}_{saliency}(o, s, f) = w_1^s \, \mathcal{L}_{BCE}(o, s) + w_2^s \, \mathcal{L}_{MSE}(o, s) + w_3^s \, \mathcal{L}_{NSS}(o, f) + w_4^s \, \mathcal{L}_{CC}(o, s), \tag{6}$$

where o is the generated saliency map and s is the ground truth saliency map which is obtained by convolving a Gaussian filter on the binary eye fixation map f for the corresponding image. w_1^s, w_2^s, w_3^s, w_4^s are the weights associated each loss term in the saliency loss function $\mathcal{L}_{NSS}$ is the inverse of normalized scanpath saliency and is given by,

$$\mathcal{L}_{NSS} = -\Big[\frac{1}{N} \sum_i \bar{o}_i \times f_i\Big]^{-1}, \tag{7}$$

where N is the total number of non-zero fixation locations in f and $\bar{o}$ is predicted saliency map normalized by it's mean and standard deviation. $\mathcal{L}_{CC}$ is the negative of cross correlation determined between o and s. $\mathcal{L}_{BCE}$ and $\mathcal{L}_{MSE}$ represent the binary cross entropy loss and the mean square error respectively which is computed between o and s. We explain our choice of saliency loss function in an ablative study done in section 5.

Finally, we combine the task-specific losses with $\mathcal{L}_{MMD}$ to handle the cross-domain MTL. Thus, the overall loss function,combining Eq.s 4, 5, 6, for our multi-task network is given by,

$$\mathcal{L} = w_1 \, \mathcal{L}_{detect} + w_2 \, \mathcal{L}_{saliency} + w_3 \, \mathcal{L}_{MMD}, \tag{8}$$

where w_1, w_2 and w_3 are the optimization constants associated with each task.

4. Implementation Details

4.1. Architectural Details

Figure 3 shows the details of the proposed architecture. We use ResNet-50 [21] as the backbone CNN as it outperforms other feature extractors (see Table 4).

Object Detection Stream: To detect objects at different scales in an input image, we add convolutional layers, that progressively decrease the size of the feature maps, in the base network similar to SSD [36]. The topmost network branch in Figure 3 depicts the architecture of the object detection stream.

Saliency Estimation Stream: The saliency estimation stream consists of Atrous Spatial Pyramid Pooling (ASPP) module [9] which consists of multiple convolutions with varying dilation factors to capture information from an image at varying scales. The ASPP module is followed by a small decoder network consisting of three convolutional and upsampling layers with ReLU activations. There is an additional 1×1 convolution layer with sigmoid activation to generate the saliency heat maps. The different loss terms are scaled to similar order of magnitudes and thus, w_1^s, w_2^s, w_3^s, w_4^s, in Eq. 6 are set to 1.0, 10.0, 1.0, 1.0 respectively. The lower network branch in Figure 3 depicts the architecture of the saliency estimation stream.

4.2. Data Preprocessing and Training

Generally, label preserving image transformations effect the human gaze annotations resulting in inaccurate saliency maps. However, some transformations, such as mirroring, inversion, shearing, compression and noise injection, do not affect the human gaze [8]. Considering this, we preprocess the training images with random horizontal and vertical flipping and Gaussian noise injection for the two tasks to improve the generalizability of the model.

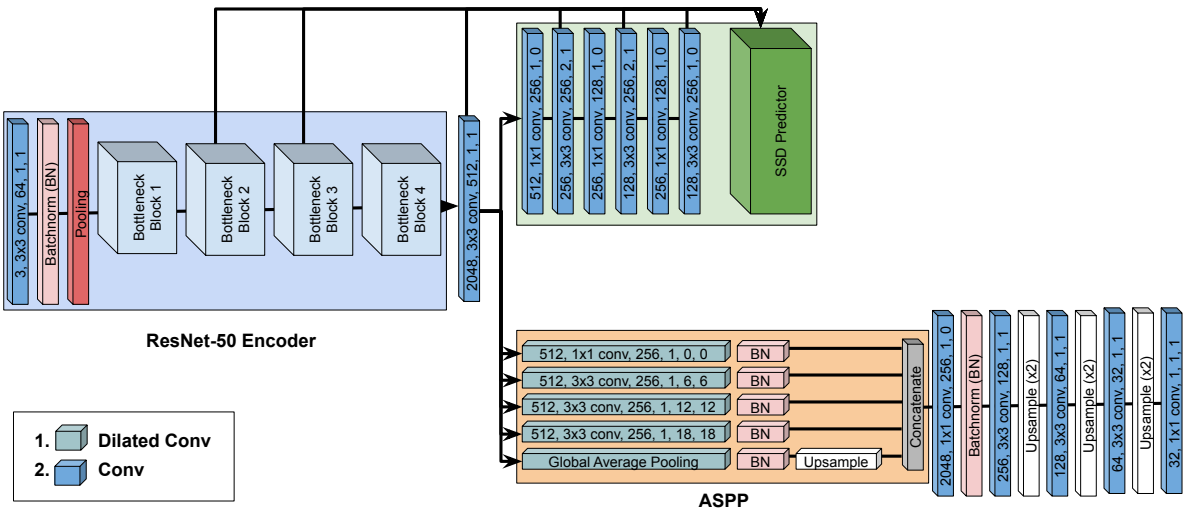

Figure 3. Detailed architecture of the proposed network. The layer hyperparameters are depicted in (in channel, kernel size, out channel, stride, padding, dilation) format. BN represents the batch normalization layer.

We train each of the task streams using only those datasets which contain annotations specific to the task. Thus, shared encoder processes two samples, which is sampled for each stream, in each iteration and learns relationship between the two source domains using the MMD loss. All the models, in all of our experiments, are trained with SGD optimizer with a learning rate of $1.0e^{-3}$, momentum 0.9 and a weight decay factor of $5.0e^{-4}$. We set the optimization parameters, w_1, w_2, w_3, in Eq. 8 to 1.0, 1.0, and 10.0 respectively. We chose these weights to ensure that different loss objectives have similar order of magnitudes.

5. Experiments And Results

5.1. Experimental Setup

We conduct our experiments on two popular datasets: MS-COCO [33] and Pascal-S [31] (which consists of images from Pascal VOC dataset) datasets. MS-COCO provides bounding box and class label annotations for object detection and Pascal-S provides the eye-fixation annotations for saliency estimation. We perform all the experiments using PyTorch [38] framework on a machine with Intel Xeon 2.4 GHz processor, 128 GB memory and Nvidia Quadro P6000 GPU with 24 GB memory.

5.2. Evaluation Criteria

We use Jaccard index between the ground truth and the predicted bounding boxes to evaluate the object detection task. We report the mAP score by thresholding the Jaccard index between 0.5 to 0.95. For saliency prediction task, we evaluate our model using the metrics - NSS, CC, SIM, KL-Div, AUC(Judd), and AUC(Borji), as defined in [26].

5.3. Baselines

We prepare two baselines: a single task network (STN) for object detection, i.e, SSD300 [36], and an STN for saliency estimation, i.e, MSI-Net [30]. We replace the VGG-16 feature extractor in both the baselines with the ResNet-50. We compare the proposed model with these baselines to show that joint training is beneficial for object detection while maintaining a competitive performance in saliency estimation. This could be credited to the domain invariant feature extractor learnt using the MMD loss. Table 1 and 2 compare the results of our method with the baselines for object detection and saliency estimation respectively.

Method	MS-COCO	
	mAP(0.5:0.95)% ↑	**mAP(0.5)% ↑**
SSD300 [36]	21.7	35.6
Proposed	**22.3**	**36.5**

Table 1. Comparison with STN for object detection on COCO 2017.

Method	Pascal-S				
	NSS ↑	**CC ↑**	**SIM ↑**	**AUC ↑**	**KD ↓**
MSI-Net [30]	**2.71**	**0.73**	**0.64**	**0.91**	**0.57**
Proposed	2.58	0.70	0.63	0.90	0.75

Table 2. Comparison with STN for saliency estimation on Pascal-S dataset.

5.4. Inference Speed and Memory Footprint

We also compare the inference speeds of our model with the STN baselines on NVIDIA Quadro P6000 GPU. The baseline networks (SSD + MSI-Net) operate at 36 FPS to jointly estimate object detection scores and saliency maps, whereas the proposed model runs at 47 FPS owing to parameter sharing in the feature extractor. In terms of memory footprint, our model occupies 901 MB for running evaluation on one image in comparison to using the two STN baselines that occupy 1035 MB, saving approximately 13% memory. We also compare on GMACs (Giga Multiply–accumulate operation per second), our model operates at 13.443 GMACs compared to using two STN baselines that have a combined 22.140 GMACs per image of size 300×300. The proposed network has 51 million parameters whereas the SSD and MSI-Net have 46 million and 39 million parameters respectively. Thus, parameter sharing in the shared encoder results in a reduction of 34 million parameters in our proposed network.

Setting	Correlation Score
Two Tasks, One Dataset	0.19
Two Tasks, Two Datasets	0.15

Table 3. Task Affinity between object detection and saliency estimation in two different settings

5.5. Task Relatedness

We use the STN baselines from Section 5.3 and report the task correlation scores. Formally, we create the Representation Dissimilarity Matrices (RDMs) [15] from the ResNet-50 features of the 100 randomly selected images from the task-specific dataset(s) of interest. We then compute a pairwise correlation between the RDMs resulting in a matrix of dimension 100×100, that is known as the RSA matrix for a specific task. The pairwise correlation is defined as, $1 - p$, where p is the Pearson's Cross Correlation [39] score between the image feature pairs. We then calculate the affinity between the RSA matrices of the two tasks using Spearman's Correlation [42]. We compute the task relatedness scores in the following settings:

1. **Two Tasks, One Dataset**: We use a SALICON dataset [25] to finetune the separate baseline models for the two tasks. SALICON dataset contains the saliency annotations for a subset of MS-COCO dataset, which contains the object detection annotations.

2. **Two Tasks, Two Datasets**: In this setting, the detection baseline is trained on MS-COCO images and the saliency baseline is trained on a separate Pascal-S dataset.

For the relatedness analysis we use STNs defined in Section 5.3 for saliency estimation and object detection tasks. We observe, from Table 3, that the object detection and the saliency estimation tasks are positively correlated and, thus, could be jointly optimized in an MTL setting. However, the task relatedness score is lower in a cross-domain setup due to the diminished correlation between the image features learnt from different domains. This makes the joint optimization more challenging.

5.6. Ablation Study

In this section, we analyse the performance of different components of our model.

5.6.1 Backbone Architecture

The choice of backbone architecture is critical in extracting information rich features which can be used by the downstream tasks. We compare the domain-invariant performance of the different CNN feature extractors in our proposed model. Table 4 summarizes the results for object detection and saliency estimation. We observe that the ResNet-50 is the most suitable choice as a feature extractor for object detection and saliency estimation in a cross-domain MTL setting. In comparison to VGG-16, ResNet-50 has a better representation power due to the increased number of layers and the residual connection.

5.6.2 Saliency Objective Function

We investigate the effect of different saliency metrics in the saliency objective function, $\mathcal{L}_{saliency}$, on the Pascal-S dataset (refer Table 5). We observe from Table 5 that an exclusive use of the cross entropy loss, which is used in classification tasks, leads to the poorest performance. This is due to the inability of the cross entropy loss to regress the saliency maps and, eventually, resulting in incorrect bright-spotted artifacts in the predicted saliency maps (further corroborated by the results in [37]). Including $\mathcal{L}_{MSE}$ and $\mathcal{L}_{NSS}$ improves the performance, however the predicted saliency maps contract towards to eye fixation locations producing dots around fixation locations. Finally, replacing $\mathcal{L}_{CC}$ with $\mathcal{L}_{NSS}$ leads to the best performance across three out of the five saliency metrics. However, we add the $\mathcal{L}_{NSS}$ term to boost the NSS scores while making a small compromise in other metrics.

5.7. Comparison and Evaluation

Figure 4 shows the object detection and the saliency map outputs of our model which is robust in diverse backgrounds, illumination, and noise in the input image. We also compare our model against different state-of-the-art methods for the two tasks and tabulated in Table 6 and Table 7. For object detection, we perform the comparison on

Backbone	Pascal-S					MS-COCO 2017		
	NSS ↑	CC ↑	SIM ↑	AUC ↑	KD ↓	mAP (0.5:0.95) ↑	mAP (0.5) ↑	mAP (0.75) ↑
VGG-16	1.96	**0.75**	**0.72**	0.87	**0.40**	20.1	35.0	19.8
EfficientNet-b3	2.07	0.55	0.55	0.89	0.82	16.6	30.1	16.5
DenseNet-161	1.87	0.52	0.54	0.89	0.91	19.6	33.6	20.0
ResNet-50 (Ours)	**2.58**	0.70	0.63	**0.90**	0.75	**22.3**	**36.5**	**23.5**

Table 4. Saliency results on Pascal-S dataset and object detection mAP scores on MS-COCO 2017 Validation set with different shared encoder as backbone.

Loss Function	Pascal-S				
	NSS ↑	CC ↑	SIM ↑	AUC ↑	KD ↓
$\mathcal{L}_{BCE}$	2.47	0.67	0.61	0.73	0.68
$\mathcal{L}_{BCE} + \mathcal{L}_{MSE} + \mathcal{L}_{NSS}$	2.67	0.73	0.62	0.90	0.60
$\mathcal{L}_{BCE} + \mathcal{L}_{MSE} + \mathcal{L}_{CC}$	2.68	**0.74**	**0.64**	**0.91**	**0.56**
$\mathcal{L}_{BCE} + \mathcal{L}_{MSE} + \mathcal{L}_{CC} + \mathcal{L}_{NSS}$	**2.71**	0.73	**0.64**	**0.91**	0.57

Table 5. Saliency evaluations for different loss functions on the Pascal-S dataset

Method	Backbone	MS-COCO		
		mAP(0.5:0.95)% ↑	mAP(0.5)% ↑	mAP (0.75) ↑
SSD300* [36]	VGG-16	20.1	35.0	19.8
SSD300*	Resnet-50	21.7	35.6	23.0
SSD300 [23]	MobileNet	19.3	-	-
Proposed	Resnet-50	**22.3**	**36.5**	**23.5**

Table 6. Comparison of the proposed model against single shot object detectors on COCO 2017, * indicates the detection scores were obtained after training the model from scratch.

Method	Pascal-S				
	NSS ↑	CC ↑	SIM ↑	AUC ↑	KD ↓
DVA [47]	2.26	0.67	0.52	0.89	0.78
SALGAN [37]	1.82	0.50	0.51	0.86	1.14
EMLNet [24]	2.32	0.62	0.57	0.89	0.77
MSI-Net [30]	**2.61**	**0.82**	**0.67**	**0.91**	**0.65**
Proposed	2.58	0.70	0.63	0.90	0.75

Table 7. Comparison of the proposed model against recent saliency estimation models on Pascal-S.

MS-COCO dataset with one shot detectors and for saliency we compare on the Pascal-S dataset. We observe that the proposed model outperforms SSD with different backbone networks in a cross-domain MTL setting owing to the improved generalizability due to the inductive bias introduced by the MTL framework. For the saliency estimation task, although our model ranks behind MSI-Net it strikes a right balance between accuracy and efficiency as the proposed model consists of $\approx 100M$ less parameters than MSI-Net.

6. Conclusions

We have introduced a novel architecture for cross domain multi-task learning for scene understanding with the specific focus on object detection and saliency estimation. Apart from being light weight due to parameter sharing, our method demonstrates better performance than the baseline STNs. We design a learnable bilinear kernel function to produce domain invariant features. The proposed framework could be extended to different complex computer vision tasks, with positive task relatedness, for a holistic scene understanding.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1597–1604. IEEE, 2009. 2

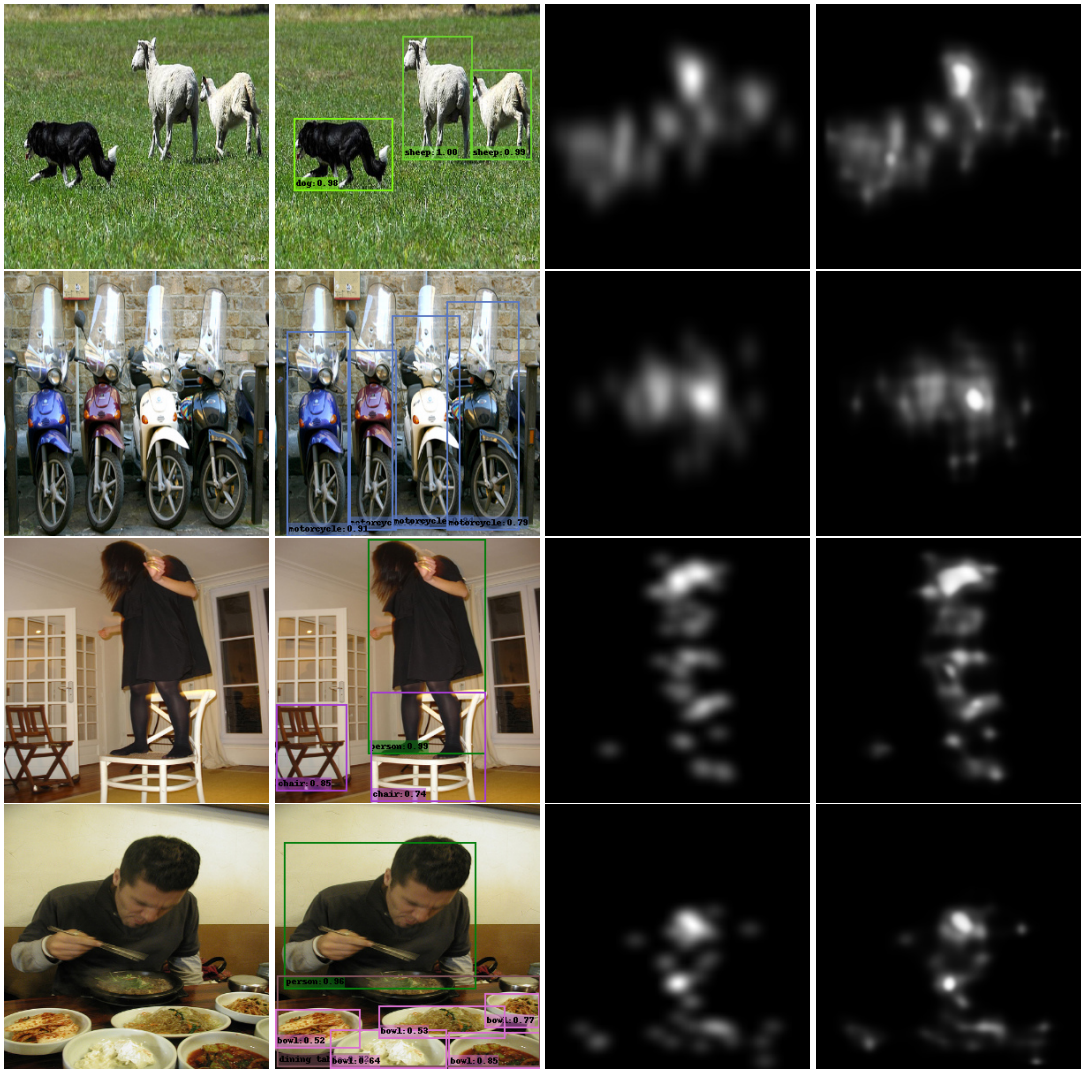

Figure 4. Qualitative results of the proposed model. From left to right, input image, detections with scores higher than 0.5, ground truth saliency map, and the predicted saliency map respectively.

[2] Shai Ben-David and Reba Schuller Borbely. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine learning*, 73(3):273–287, 2008. 1

[3] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017. 1, 4

[4] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006. 3

[5] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. 2015. 2

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020. 2

[7] Richard Caruana. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Morgan Kaufmann, 1993. 1, 2

[8] Zhaohui Che, Ali Borji, Guangtao Zhai, Xiongkuo Min, Guodong Guo, and Patrick Le Callet. How is gaze influenced by image transformations? dataset and model. *IEEE Transactions on Image Processing*, 29:2287–2300, 2019. 2, 4

[9] 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. 3, 4

[10] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, pages 794–803. PMLR, 2018. 2

[11] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018. 2

[12] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, 2016. 2

[13] Wenhui Diao, Xian Sun, Xinwei Zheng, Fangzheng Dou, Hongqi Wang, and Kun Fu. Efficient saliency-based object detection in remote sensing images using deep belief networks. *IEEE Geoscience and Remote Sensing Letters*, 13(2):137–141, 2016. 1

[14] Nikita Dvornik, Konstantin Shmelkov, Julien Mairal, and Cordelia Schmid. Blitznet: A real-time deep network for scene understanding. In *Proceedings of the IEEE international conference on computer vision*, pages 4154–4162, 2017. 2

[15] Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12387–12396, 2019. 1, 6

[16] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017. 2

[17] 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. 2

[18] 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. 2

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[20] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. 2007. 2

[21] 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. 1, 4

[22] Srinidhi Hegde, Jitender Maurya, Aniruddha Kalkar, and Ramya Hebbalaguppe. Smartoverlays: A visual saliency driven label placement for intelligent human-computer interfaces. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 1

[23] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 7

[24] Sen Jia and Neil DB Bruce. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, page 103887, 2020. 7

[25] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 6

[26] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009. 5

[27] Ramakrishna Kakarala and Ramya Hebbalaguppe. A method for fusing a pair of images in the jpeg domain. *Journal of real-time image processing*, 9(2):347–357, 2014. 1

[28] 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. 2

[29] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865. PMLR, 2017. 1, 2

[30] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020. 1, 5, 7

[31] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, 2014. 5

[32] Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Deng, and Jian Sun. Light-head r-cnn: In defense of two-stage object detector. *arXiv preprint arXiv:1711.07264*, 2017. 2

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 5

[34] An-An Liu, Ning Xu, Wei-Zhi Nie, Yu-Ting Su, and Yong-Dong Zhang. Multi-domain and multi-task learning for human action recognition. *IEEE Transactions on Image Processing*, 28(2):853–867, 2018. 3

[35] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):353–367, 2011. 1

[36] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European con-*

ference on computer vision, pages 21–37. Springer, 2016. 1, 2, 3, 4, 5, 7

[37] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017. 2, 6, 7

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc., 2019. 5

[39] Karl Pearson. Correlation coefficient. In *Royal Society Proceedings*, volume 58, page 214, 1895. 6

[40] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2

[41] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 762–771, 2018. 3

[42] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471, 1987. 6

[43] Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000. 3

[44] 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. 2

[45] 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. 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[47] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2017. 7

[48] Wenguan Wang, Jianbing Shen, Xingping Dong, and Ali Borji. Salient object detection driven by fixation prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1711–1720, 2018. 1

[49] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 1

[50] 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. 2

[51] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 2

[52] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 714–722, 2018. 2

[53] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4106–4115, 2019. 2