Reducing Distributional Uncertainty by Mutual Information Maximisation and Transferable Feature Learning

Jian Gao[1,2], Yang Hua[1], Guosheng Hu[2,1], Chi Wang[1,2], and Neil M. Robertson[1]

[1] EEECS/ECIT, Queen's University Belfast, UK
{jgao05, y.hua, cwang38, n.robertson}@qub.ac.uk
[2] Anyvision, Belfast, UK
huguosheng100@gmail.com

Abstract. Distributional uncertainty exists broadly in many real-world applications, one of which in the form of domain discrepancy. Yet in the existing literature, the mathematical definition of it is missing. In this paper, we propose to formulate the distributional uncertainty both between the source(s) and target domain(s) and within each domain using mutual information. Further, to reduce distributional uncertainty (e.g. domain discrepancy), we (1) maximise the mutual information between source and target domains and (2) propose a transferable feature learning scheme, balancing two complementary and discriminative feature learning processes (general texture learning and self-supervised transferable shape learning) according to the uncertainty. We conduct extensive experiments on both domain adaption and domain generalisation using challenging common benchmarks: Office-Home and DomainNet. Results show the great effectiveness of the proposed method and its superiority over the state-of-the-art methods.

Keywords: Distributional uncertainty, domain discrepancy, mutual information, object shape, self-supervised learning

1 Introduction

A fundamental assumption in machine learning is the similarity of training and test distribution. Various algorithms have been proposed based on this assumption, Convolutional Neural Networks (CNNs) among which achieved huge success together with large scale training data. However, opposed to this ideal setting, distributional uncertainty exists broadly in almost every real-world problem.

Domain discrepancy is a type of distributional uncertainty in which the dissimilarity of two domains (distributions) is considered. Unsupervised Domain Adaptation (UDA) aims at resolving the domain discrepancy problem and enhancing model transferability, while not requiring any labels for target domain. Different approaches have been proposed to tackle it, such as direct minimisation of domain discrepancy [32] and domain adversarial learning [15]. Though promising progress has been made, some critical issues still remain.

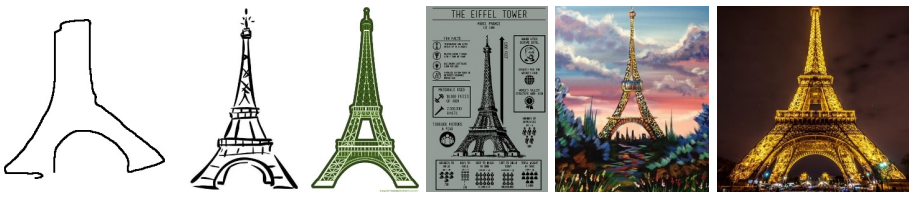


Fig. 1. Examples of the Effiel Tower presented in different styles from the six domains in DomainNet dataset [38]. Left to right: quickdraw, sketch, clipart, infograph, painting and real. *Best viewed in colour.*

First of all, there lacks a unified and quantified explanation for the concept of domain discrepancy. Existing methods either minimise the distributional difference between classifier outputs [44], or try to align intermediate features [49] from different distributions. Yet a rigorous definition of domain discrepancy and a precise measurement of it are missing.

Secondly, most methods are restricted to aligning a single source and a single target domain at a time, as they assume that training and test data each follows a single distribution. However, this is often not true in practice. In some tasks, the presence of multiple distinct distributions in training is almost unavoidable, for example, different camera-angle sub-domains in person re-identification tasks [40]. Directly applying a single source to single target adaptation method to such tasks is problematic. Since the adaptation performance across multiple domains is bounded by the worst model obtained from a source domain that is least similar to the target [56].

Further, contemporary methods try to align the output distributions by one or more classifiers [35], while leaving the feature learning entirely handled by CNNs. However one main drawback of CNNs is its lack of regularisation in learning generalised and transferable features [39]. For example, CNNs are heavily biased towards learning textures which may change dramatically across domains, while neglecting object structural features such as shape that is often more consistent [17]. For example in Figure 1, the Effiel Tower appears in visually diametrically different image styles but its shape remains consistent. The situation becomes even worse when a distributional dissimilarity lies between the training and the test data [38].

From our observations, the above issues are inherently caused by the same fact which is the distributional uncertainty. The domain discrepancy in UDA describes the distributional uncertainty *between source and target*. Single source adaptation methods fail to consider the distributional uncertainty *within the source samples*. The lack of regularisation in CNNs can be compensated by a reduction in the distributional uncertainty of training data, such as providing certain prior knowledge about the distribution.

In this paper, we propose to resolve the above issues by reducing distributional uncertainty, combing Mutual Information Maximisation and Transferable Feature Learning (abbreviated as **MIMTFL**). We formulate the estimation of distributional uncertainty using Mutual Information (MI). During training,

we calculate MI over each batch of samples to exam its uncertainty. We learn to reduce the uncertainty by maximising MI between source and target, while considering uncertainty within the source samples. We further leverage a self-supervised transferable feature learning scheme by enforcing a balance between texture and object shape features. According to the estimated uncertainty level, the network learns to automatically balance texture and shape features for better transferability.

In summary, we propose the following contributions:

- A formulation to measure distributional uncertainty as a unified definition of domain discrepancy. The proposed distributional uncertainty measurement using mutual information is mathematically grounded, and generalise to not only discrepancy between different domains, but also disagreement within source.
- A self-supervised transferable feature learning strategy that utilises MI to automatically balance the learning of texture and shape features for better generalisation.
- Extensive results under various settings on two large-scale multi-domain adaptation benchmarks with state-of-the-art performance to prove the effectiveness of the proposed method.

2 Related Work

2.1 Unsupervised Domain Adaptation

The main challenges in UDA come from two aspects. First, how to make the source model transferable to the target in view of the distributional gap between the two. Secondly, how to make use of the unlabelled target samples. For the first, learning a transferable representation for both the source and target has attracted much attention. Methods including DDC [50], DAN [30] and JAN [33] align the domains by minimising a domain discrepancy measurement between them. Domain adversarial learning [15,49] seems to be effective where the gradients from a domain discriminator network trying to distinguish source and target samples are reversed. It is also found that aligning the first, second and even higher order moments of source and target distributions is helpful [47,43].

To use the unlabelled target samples, many semi-supervised learning techniques are introduced into UDA algorithms. Pseudo-Labelling [28] and Label-Propagation [58] are found to be useful to estimate the true labels of target samples [5]. Another effective solution is the Mean-Teacher [48] model in which an unsupervised consistency loss is enforced between a student model prediction and a teacher model prediction [14].

2.2 Distributional Uncertainty

The study of distributional uncertainty spreads through various fields of science and engineering. In control theory, researchers apply distributional uncertainty

analysis to enhance robustness of controllers [34]. Evidence imprecision and uncertainty modelling using fuzzy sets is shown critical in medical diagnosis [46]. Probabilistic machine learning values the representation and manipulation of uncertainty in both data distributions and models, as it plays a central role in scientific data analysis [18].

The connection between distributional uncertainty estimation and generalisation has recently been uplifted significantly too. In modern deep learning, explicit modelling of distributional uncertainty within the Bayesian framework makes the network more robust to noisy data as well as achieving better generalisation results on difficult computer vision tasks, such as semantic segmentation and depth regression [24]. Kendall et al. [25] found that using uncertainty weighting in multi-task learning allows effective simultaneous learning of various tasks, which even outperforms individually learned models for each task. Uncertainty-based reliability analysis in 3D vehicle detection from point cloud data brings steady improvement in the model's performance in adverse environment, such as heavy occlusions, which is of great importance to promoting safe autonomous driving [13]. Examples of benefits in generalisation can be found throughout a variety of vision applications including optical flow estimation [22], people tracking in traffic scenes [2], so on and so forth.

2.3 Self-Supervised Visual Feature Learning

As a promising solution towards eliminating the need of costly human annotations, self-supervised learning methods learn visual features from unlabelled images on auxiliary tasks. The supervision signals for the auxiliary tasks are usually automatically obtained without requiring any human labelling effort. In other words, it absorbs advantages from supervised learning methods with accurate labels, while saving the tedious labour for labelling.

So far, outstanding progress has been made in a wide range of vision tasks with self-supervised visual feature learning. Zhang et al. [55] found learning to colourise of de-coloured images eminently improve generalisation in the downstream recognition task. With an image inpainting auxiliary task, the network captures not just appearance but also the semantics of visual structures [37]. Another popular method to learn structural features in an image is extracting patches from it, and learning to predict their relative spatial locations, so-called solving a jigsaw puzzle. The learned visual representation is seen to perform incredibly well in object detection tasks [10], as well as effectively increasing generalisation in standard Domain Generalisation tasks [4].

3 Methodology

We hereby introduce the details of **MIMTFL**, as illustrated in Figure 2. We first elucidate the measurement of distributional uncertainty formulation using MI in §3.1. Then move on to the MI-guided self-supervised learning of transferable features including our proposed shape learning method, detailed in §3.2.

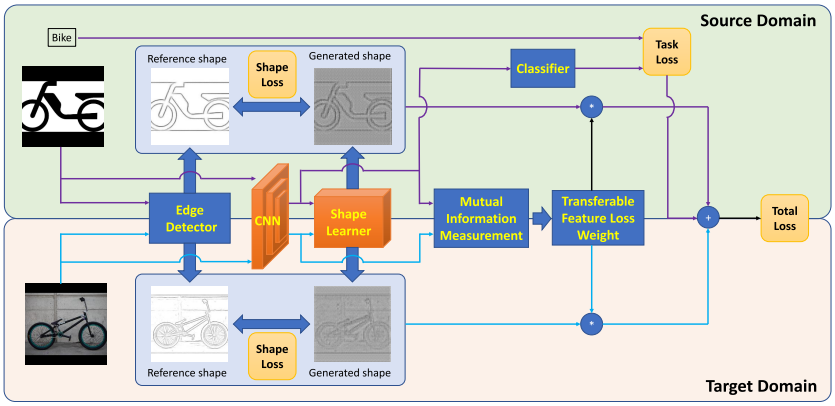


Fig. 2. System pipeline: A reference shape is obtained for all the samples using an edge detector in advance. During training, both the source and the target sample batch are fed to the backbone feature extractor CNN. A shape loss is calculated via comparing the generated shape by the shape learner and reference shape for each sample. The CNN extracted features are also fed into a mutual information measurement module, to estimate both inter-domain and intra-domain distributional uncertainty for source and target domain. The estimated uncertainty is then used to automatically weight the shape loss. The overall objective consists of the classifier task loss on the labelled source samples, and the weighted shape loss on all the samples. *Best viewed in colour.*

3.1 Distributional Uncertainty Measurement using MI

Preliminaries. For distributional uncertainty estimation, we are interested in the problem: given two batches of samples, how can we know whether they follow the same distribution? Specifically, how can we measure the similarity between them? In information theory, the measure of uncertainty on a distribution $p(X)$ is the entropy H of X [45] on the sample space $\mathcal{X}$, given as:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -\mathbb{E}[\log p(X)]. \tag{1}$$

Uncertainty Measurement. Mutual Information measures the reduction in uncertainty for one variable X given a known value of another variable Y, which is defined by:

$$I(X;Y) = H(X) - H(X|Y). \tag{2}$$

MI between X and Y can be calculated using the Kullback-Leibler(KL) divergence [26] between the joint entropy $H(X,Y)$ and product of the two individual entropy $H(X)H(Y)$.

$$I(X;Y) = D_{KL}(H(X,Y)||H(X)H(Y)). \tag{3}$$

We calculate $I(X;Y)$ between two batches of samples X and Y as the confidence of them belonging to the same distribution using Eq. (3). Note that for a single distribution $p(X)$, ideally, the MI within its observations is the entropy itself:

$I(X;X) = H(X)$. However, we need to re-evaluate the value to detect noise in the samples, which indeed often exist in real-world. Therefore, to measure the uncertainty within a distribution, we use the below formula:

$$I(X;X) = D_{KL}(H(X,X)||H(X)H(X)). \tag{4}$$

In practice, the higher the measured MI, the lower the distributional uncertainty. During learning, we maximise the MI between source and target to reduce the uncertainty and learn a more generalised model. Note that although there is also possibility that the MI between distributions [1] can be learnt instead of calculated, it cannot be directly applied here. Since this would require the distributions to be known, while we have one of the distributions (the target) unknown.

3.2 Transferable Feature Learning

MI-Guided Transferable Feature Learning. Geirhos et al. [17] found that CNNs recognise objects mainly according to their texture, while overlooking other structural features such as shape and edges. This texture bias is observed repeatedly with different network architectures across different tasks [53,41]. To increase model transferability, we need to reduce such bias by introducing appropriately balanced learning of texture and other non-texture features. In tasks where texture varies drastically across observations, learning of other transferable features should be attenuated. Whilst if texture may serve as a generic feature across different distributions, the learning of texture should not be diminished.

We propose to further utilise the distributional uncertainty measured by mutual information in §3.1 as the controlling factor for transferable feature learning. To be specific, we adopt the below learning object:

$$\mathcal{L}_{total} = \mathcal{L}_{cls}(X) + \lambda_X \cdot \mathcal{L}_{trans}(X) + \lambda_Y \cdot \mathcal{L}_{trans}(Y), \tag{5}$$

$$\lambda_X = \frac{I(X;X)}{I(X;Y)}, \lambda_Y = \frac{I(Y;Y)}{I(Y;X)}, \tag{6}$$

where $\mathcal{L}_{cls}$ is the supervised classification loss on the labelled source samples X and $\mathcal{L}_{trans}$ the auxiliary loss for explicit learning of non-texture transferable features on both X and unlabelled target samples Y.

The weighting term λ_X and λ_Y, decided by MI, dedicates to balance between texture and non-texture feature learning. The intuition is that, when MI is small which indicates high distributional uncertainty, learning of more transferable features can be beneficial. Whilst larger MI manifests a successful transfer within the network and thus alleviating the need for complementary features other than texture.

The selection of a proper $\mathcal{L}_{trans}$ can be exhaustive. One option is increasing the texture diversity in the training data. For example, some involve comprehensive pre-defined image augmentations [54,9,7], some apply learned style transfer

[17,29]. However, the improvement in their generalisation comes at the cost of huge computations in retrieving the enormous potential sample space. Not to mention that, diverse training samples are often required in the first place to learn appropriate augmentation/transformation methods.

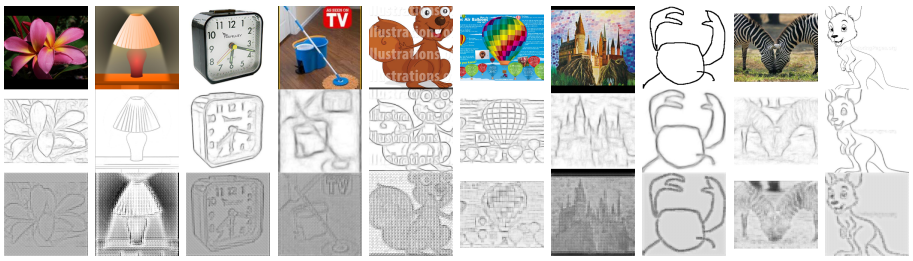
Self-Supervised Visual Feature Learning. Since the requirement for exhaustive annotations has been identified as a major bottleneck for deep learning, the concept of self-supervised learning is proposed as a promising solution [10]. In self-supervised learning, rich information in the input that may be ignored by the designated task learner is further mined by an auxiliary task. And the auxiliary task can be trained without requiring any human labelling effort. Examples of such tasks are colourisation of de-coloured images [55,27], jigsaw puzzle [10,4] and inpainting [37]. Empirical results using these auxiliary tasks are found helpful in regularising the learning procedure and notably improving generalisation [20]. Hence, we propose to employ self-supervised learning methods to explicitly enforce non-texture feature learning.

One thing to note for our selection of the auxiliary loss $\mathcal{L}_{trans}$ is that, no labels should be required for the unlabelled target samples Y. And this fits seamlessly into the setting of self-supervised learning. In theory, the aforementioned tasks should be compatible within our framework, where their corresponding learning objectives can be the choice of $\mathcal{L}_{trans}$ in our MI-guided transferable feature learning.

While most of these methods are proven effective to regularise the network learning to be less biased towards easy-to-fit texture, none of them exploits object shape explicitly. This is, however, contradictory to the way by which human recognise visual objects. Neuroscience study [19] found that, training human participants on recognising objects composed of certain object shapes significantly improves their performance in recognising other different objects containing the same shapes. Motivated by the fact that shape plays a vital role in human's visual perception and recognition paradigm, we propose a new self-supervised shape learning task to better mimic human vision using CNNs. Note that our new shape learning method serves as one new potential candidate for the choice of $\mathcal{L}_{trans}$, and is in parallel with the aforementioned ones.

Self-Supervised Object Shape Learning. The target of object shape learning is to embed object structural information into the un-constrained features that the network learns. Learning object shape as a complementary feature is advantageous not only because it is more interpretable. In the existence of sharp change in texture and colour, recognition using object shape is more reliable as it is often more consistent and independent of the frequent appearance variations.

Based on the principle that we would like to avoid requiring any extra labelling effort in creating a reference shape for each training sample, we design a self-supervised shape learning scheme by creating reference shapes using edge detectors. Edge detection as a traditional low-level vision task has been studied thoroughly with mature tools formulated. The overall object shape can be obtained by running an edge detector on an image. Here, the target of our learning is a set of shape-embedded generic features that can contribute to the visual



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 3. Examples in Office-Home (a-d) and DomainNet dataset (e-h): 1^{st} row - original, 2^{nd} row - reference shape and 3^{rd} row - generated shape image trio.

recognition end. Motivated by the Perceptual Loss [23], we define a feature descriptor f that captures the low-level feature of an image. A shape learner learns to generate a shape image given an input object image. Examples of shapes extracted by traditional edge detector and generated shape by our framework are illustrated in Figure 3. Then we measure the perceptual difference between low-level features of the generated shape f_{gen} and the reference shape f_{ref} as:

$$\mathcal{L}_{shape} = \mathcal{D}(f_{gen}, f_{ref}), \tag{7}$$

where $\mathcal{D}$ is a deviation measure for the low-level features.

Compared to pixel-level loss, our shape loss gives more freedom to the shape learner to keep certain degrees of texture-related details that can benefit the recognition process in some cases (such as the shade in Figure 3 (b)). On the contrary, strictly constraining the shape learner to reference shape at the pixel-level obviously would lose this discriminative information.

4 Experiments

In this section, we provide a detailed empirical analysis of MIMTFL. We conduct extensive experiments on two large scale datasets with varying distributional uncertainties. To analyse the effectiveness of MIMTFL, we consider these three aspects: i) the level of distributional uncertainty exists in these existing evaluation benchmarks; ii) the benefit of uncertainty reduction and iii) the benefit of explicit shape learning.

4.1 Settings

Unsupervised Domain Adaptation (UDA). In UDA experiments, the training set consists of two types of data: labelled samples from one or more distributions (source), and unlabelled samples from one or more distributions (target). We further divide these tasks into three groups, which correspond to most common real-world scenarios, with the difficulty level in each group magnified compared to the previous: i) **Single-to-single** adaptation in which we train

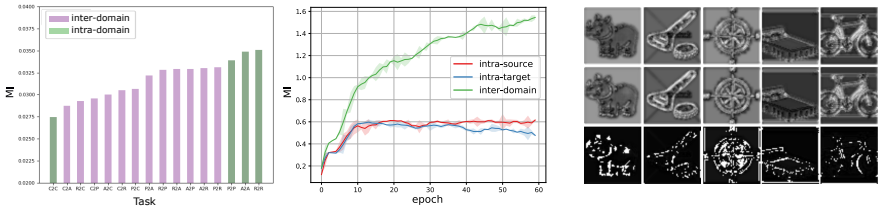


Fig. 4. Measured MI during training in Office-Home experiments, the higher the MI, the lower the distributional uncertainty. (a) Initial MI both intra and inter-domain in Office-Home Dataset, (b) the change of MI during training in R→P experiment, (c) visualised features without (1^{st} row) and with (2^{nd} row) shape learning, and the difference between them (3^{rd} row). *Best viewed in colour.*

using a single labelled source distribution and a single unlabelled target distribution. We evaluate the effect of our self-supervised shape feature learning on model transferability with limited samples, and the generalisation capacity of our model across these different tasks. ii) **Multi-to-single** adaptation where the labelled source samples come from multiple distributions, whilst the unlabelled target samples follow a single distribution. We validate the benefits of our framework to simultaneous learning from multiple disjoint distributions. iii) **Multi-to-multi** adaptation with both the labelled source and unlabelled target are drawn from multiple distributions. We verify the effect of our proposed method with up-scaled unlabelled target data and even fewer labelled source data.

Domain Generalisation (DG). DG is a more challenging task since the target is completely absent during training. Under this setting, we can only estimate the distributional shift within the labelled source domain, yet still transfer the learned model to the target at test time. To achieve a reasonable result on target, a model has to learn as much as possible the most transferable and generalised features from the given source distributions.

4.2 Datasets

Office-Home. The dataset [51] parents four distinct domains in 65 object categories: Art (A), Clipart (C), Product (P) and RealWorld (R). The number of images in each domain is: 2427, 4365, 4439 and 4357, respectively. For all the experiments, we use all the samples in each domain as either source or target. Note that, although we treat each domain as a single distribution following common practice, there still exist varying levels of out-of-distribution examples in some domains. For instance, the Clipart domain which contains artistic images in various forms and patterns exhibits a rather high distributional uncertainty even within the domain itself. This is clearly indicated in its low measured MI value shown in Figure 4 (a).

DomainNet. As the largest and the most challenging benchmarking dataset for DA till today, DomainNet dataset [38] includes six domains: clipart (c),

infograph (i), painting (p), quickdraw (q), real (r) and sketch (s). The samples spread through 345 diverse categories and each domain is divided into a train split for training and a test split for test only. The difficulty of this dataset comes from several aspects: i) highly abstract and artistic images, ii) diverse domains, and iii) severely unbalanced domains.

4.3 Implementation Details

Domain adaptation. We apply MIMTFL on both Office-Home and DomainNet for all the three types of tasks. In single-to-single adaptation experiments, we use each domain as source and another domain as the target in turn for all the domains, ending up with 12 single-to-single tasks for Office-Home and 30 single-to-single tasks for DomainNet. Specifically, for better evaluation of different levels of distributional uncertainty, we divide 30 single-to-single adaptation tasks in DomainNet into two groups: 15 difficult tasks with average target accuracy by the source only model at 10.8% and 15 easier tasks at 42.3%. We report the mean accuracy separately for the two groups and an all mean accuracy for all of the 30 tasks. In multi-to-single adaptation experiments, target is only one domain while the source comes from multiple different domains. We use each domain as the target and the rest as source in turn, resulting in four tasks for Office-Home and six for DomainNet. The combined scale of the source domain is usually much larger than the target. In multi-to-multi experiments, we use all four domains in Office-Home and four domains in DomainNet that yield the highest distributional gap: infograph, quickdraw, real and sketch. In each task, two domains are chosen as source and the other two as target.
Domain generalisation. To compare with existing domain generalisation methods, we conduct single-to-single and multi-to-single experiments on Office-Home dataset. Under this setting the shape loss is only applied to the source samples, where no target samples are used in training.
Reference shape extraction. While many well-developed edge detectors including Canny [3] are generally applicable, we use a structure-aware edge detector [11] here as an example. Each image is input into the algorithm, and the output is a single channel image under the same resolution of the input. Object shapes are described by the edges, resulting in a sketch-lized representation of the original image. For most images, the extracted edges well depict the object shapes. For all our experiments, we use the Structured Forests edge detector implemented in OpenCV[3].
Network and training setups. We use the original ResNet [21] architecture with a single FC layer as our backbone network for all experiments. For the relatively smaller Office-Home dataset, we use ResNet-50 for UDA experiments and ResNet-18 for DG experiments. For DomainNet experiments, we use ResNet-101 backbone. Following common practice, ImageNet [8] pre-trained model is used as backbone network initialisation. The shape loss compares shape features extracted from the first convolutional block output in ResNet. The weight of

[3] https://opencv.org

the shape loss is calculated using the discrepancy between output features of the first residual block output in ResNet. The shape learner is composed of two layers of deconvolution, with output channels 128 and 3, respectively. The generated shape image is exactly the same dimension with the input image, fixed as $224 \times 224 \times 3$. We use SGD [42] optimiser with learning rate 0.0001 for the ImageNet pre-trained layers and 0.001 for the FC layer in all the experiments. We apply basic data augmentation routine including random cropping and horizontal flipping only. All implementations and experiments are based on PyTorch [36].

Benchmarking. We compare our method with representative and latest state-of-the-art algorithms that study the problem of distributional discrepancy. Following the UDA setting, SE [14], CDAN [31] and DWT [43] are the latest top-performing DA methods on the Office-Home dataset. DANN [16] and ADDA [49] are representative domain adversarial learning methods, while JAN [33] and MCD [44] focus on discrepancy minimisation. For a fair evaluation in Office-Home experiments, we also report result using our method combined with the Min-Entropy Consensus loss (MEC) [43] to compare with aggregated methods: DWT+MEC [43] and BSP+CDAN [6]. For DG setting, JiGen[4] uses self-supervised task of solving jigsaw puzzles to increase model transferability and is the state-of-art on Office-Home dataset. D-SAMs [12] is a representative method that targets at bridging multiple distinct distributions in training for DG. All the results are reported in top 1 accuracy and cited from the original papers.

4.4 Ablation Study

Uncertainty analysis in existing datasets. We compare the average uncertainty, indicated by MI both between different domains and within the same domain illustrated in Figure 4 (a). We observe clearly that, in Office-Home dataset, the visually most diverse domain - Clipart yields the lowest MI even within itself. This is consistent with the adaptation results, where we see that tasks involving the Clipart domain are generally more difficult than others.

Component analysis of MIMTFL. To further understand the effectiveness of MIMTFL, we conduct ablation studies on its two main components, namely **MIM** for MI maximisation and **TFL** for transferable feature learning. Figure 4 (c) clearly shows that with the shape learning objective, the network focuses more on object shape rather than textures. From the results in Table 1, we observe that without either module, the adaptation results on almost all the tasks drop significantly. Especially on those difficult tasks where the source only model accuracy is lower, such as P→C and A→C, adding in the MI module increases model transferability by a large margin. The change of measured MI during training is plotted in Figure 4 (b), which shows a steady increase of the mutual information between domains, indicating the reduction in the uncertainty. Specifically, we observe that the value of the MI yields a higher impact on those more difficult adaptation tasks, where the distributional uncertainty is higher. Our MI maximisation gives the best performance on these challenging tasks. As in reality, it is difficult to manually tune the hyperparameter for transferable

Table 1. Results on Office-Home single-to-single UDA (all with ResNet-50 backbone).

Method	P→C	A→C	C→A	P→A	R→C	C→P	C→R	A→P	R→A	A→R	P→R	R→P	mean
Source only	31.2	34.9	37.4	38.5	41.2	41.9	46.2	50.0	53.9	58.0	60.4	59.9	46.1
DDAIG [57]	36.8	40.8	43.7	40.0	44.5	55.6	56.9	53.6	56.3	65.4	63.5	73.8	52.6
MIMTFL$^*_{src}$	**48.5**	**51.5**	**57.1**	**53.2**	**52.2**	**65.6**	**67.6**	**67.9**	**66.2**	**74.8**	**78.8**	**74.1**	**63.1**
DANN [16]	43.7	45.6	47.0	46.1	51.8	58.5	60.9	59.3	63.2	70.1	68.5	76.8	57.6
JAN [33]	43.4	45.9	50.4	45.8	52.4	59.7	61.0	61.2	63.9	68.9	70.3	76.8	58.3
SE [14]	41.5	43.2	55.0	50.4	49.5	59.0	64.5	60.2	64.9	70.4	68.9	75.2	58.6
CDAN [31]	49.1	50.6	55.7	51.8	**56.9**	62.7	64.2	65.9	68.2	73.4	74.5	80.7	62.8
DWT [43]	49.5	50.8	58.9	**57.2**	55.3	65.6	60.2	**72.0**	**70.1**	**75.8**	**78.3**	78.2	64.3
TFL	48.6	52.5	57.2	54.8	51.4	66.2	68.6	68.4	66.5	75.0	74.0	78.2	63.4
MIM	48.8	52.1	57.1	52.8	52.6	66.0	68.5	67.5	65.9	74.2	73.5	78.1	63.1
MIMTFL	**51.1**	**54.1**	**59.1**	55.8	55.4	**66.9**	**69.4**	68.4	67.8	75.2	74.6	**79.1**	**64.7**
BSP+DANN [6]	49.6	51.4	56.0	57.0	57.1	67.8	68.8	68.3	70.4	75.9	75.8	80.6	64.9
DWT+MEC [43]	47.9	54.7	56.9	54.8	54.9	68.5	69.8	**72.3**	68.6	**77.2**	**78.1**	81.2	65.4
BSP+CDAN [6]	50.2	52.0	58.0	**58.6**	59.3	**70.3**	70.2	68.6	**72.2**	76.1	77.6	**81.9**	66.3
MIMTFL+MEC	**54.9**	**56.9**	**61.2**	**58.6**	**59.4**	70.0	**71.6**	70.3	69.8	75.6	77.5	80.4	**67.2**

Table 2. Results on DomainNet single-to-single UDA (all with ResNet-101 backbone).

Method	q→i	q→p	i→q	q→r	p→q	r→q	q→c	q→s	s→q	c→q	s→i	p→i	c→i	r→i	i→s	mean	
Source only	**0.9**	1.4	**3.6**	4.1	**4.9**	**6.4**	**7.0**	**8.3**	**10.9**	**11.1**	**15.4**	**18.7**	**19.3**	**22.2**	**27.9**	**10.8**	
MIMTFL$_{src}$	0.8	**4.8**	3.2	**4.3**	4.8	6.0	6.7	7.6	10.8	10.2	13.9	16.5	15.4	19.1	27.2	10.1	
ADDA [49]	2.6	5.4	3.2	9.9	8.4	**12.1**	15.7	11.9	**14.9**	3.2	8.9	9.5	11.2	14.5	14.6	9.7	
MCD [44]	3.0	**7.0**	1.5	11.5	1.9	2.2	15.0	10.2	3.8	1.6	13.7	14.8	14.2	19.6	18.0	9.2	
DANN [16]	2.0	4.4	**3.8**	9.8	5.5	6.3	11.8	8.4	10.4	9.5	13.9	15.1	15.5	17.9	25.7	10.7	
TFL	1.5	3.6	2.3	9.0	3.1	5.1	14.5	8.0	11.1	9.9	15.4	16.2	15.9	18.9	27.8	10.8	
MIMTFL	**3.1**	5.0	2.9	**16.0**	4.2	5.8	**18.8**	**13.8**	12.3	10.7	**16.5**	14.7	15.1	19.0	**31.0**	**12.6**	
Method	i→c	i→p	p→s	s→p	c→p	r→s	p→c	c→s	i→r	s→c	s→r	r→c	r→p	c→r	p→r	mean	all mean
Source only	30.2	**31.2**	36.3	**37.0**	**37.5**	**38.8**	39.6	41.0	44.0	46.9	47.0	48.4	**49.4**	**52.2**	54.5	42.3	**26.6**
MIMTFL$_{src}$	**33.1**	30.8	**36.8**	35.1	36.0	37.9	**41.8**	**42.3**	**46.4**	**49.4**	**47.2**	**50.0**	47.4	51.7	**56.4**	**42.8**	26.3
ADDA [49]	19.1	16.4	25.4	25.2	24.1	25.7	31.2	30.7	26.9	35.3	37.6	39.5	29.1	41.9	39.1	29.8	19.8
MCD [44]	23.6	21.2	28.4	27.6	26.1	29.3	34.4	33.8	36.7	41.2	34.8	42.6	42.6	45.0	50.5	34.5	21.9
DANN [16]	31.8	30.2	35.1	34.5	34.8	37.3	39.6	41.4	44.8	47.9	46.8	47.5	47.0	50.8	54.6	41.6	26.1
TFL	**33.7**	**32.0**	**36.8**	39.9	36.0	37.1	**43.2**	41.9	46.9	**52.2**	50.9	**50.0**	47.5	**52.4**	**56.5**	**43.8**	27.3
MIMTFL	**32.1**	31.0	**36.8**	**40.3**	35.6	**39.4**	**40.1**	**43.1**	**48.5**	**51.7**	**53.5**	**48.5**	47.6	51.5	**55.4**	**43.7**	**28.1**

feature learning, which is a strong demonstration of the benefit of our adaptive weighting according to calculated distributional uncertainty.

4.5 Results

Unsupervised domain adaptation. The results are shown in Table 1 and 2. We list the task column in the ascending order of their source only model performance. Although the Office-Home benchmark is highly competitive, it is observed that MIMTFL is able to achieve top performance. Specifically, our method achieves significant improvement on those tasks with lower source only accuracy. Such as A→C task, it outperforms previous SOTA with a 3.3% absolute gain. When combined with the MEC loss, it further boosts the average target accuracy and produces the best result among all methods in 6 out of all the 12 tasks. It is worth noting that, DWT and DWT+MEC engage affine

Table 3. Results on Office-Home multi-to-single UDA (all with ResNet-50 backbone).

Method	ACP→R	ACR→P	APR→C	CPR→A	mean
Source only	81.7	80.1	58.5	69.4	72.4
SE [14]	79.2	76.3	54.3	68.8	69.7
DWT+MEC [43]	**83.8**	**83.9**	59.1	**73.0**	74.9
MIMTFL+MEC	83.1	81.9	**64.3**	72.6	**75.5**

Table 4. Results on DomainNet multi-to-single UDA (all with ResNet-101 backbone). * indicates multi-source adaptation methods

Method	ipqrs→c	cpqrs→i	ciqrs→p	ciprs→q	cipqs→r	cipqr→s	mean
Source only	39.6	8.2	33.9	11.8	41.6	23.1	26.4
SE [14]	24.7	3.9	12.7	7.1	22.8	9.1	13.4
ADDA [49]	47.5	11.4	36.7	14.7	49.1	33.5	32.2
MCD [44]	54.3	22.1	45.7	7.6	58.4	43.5	38.6
DANN [16]	58.1	21.0	51.1	10.3	66.2	49.3	42.7
DCTN* [52]	48.6	23.5	48.8	7.2	53.5	47.3	38.2
M^3SDA* [38]	57.2	24.2	51.6	5.2	61.6	49.6	41.6
MIMTFL	**67.2**	**25.0**	**54.4**	**13.4**	**67.0**	**54.1**	**46.8**

transformation and Gaussian blurring as additional data augmentation during training, while our reported results are without such sophisticated augmentation. Additionally, we report a source only result where no target data is used in training, namely MIMTFL$_{src}$. Compared with latest method such as DDAIG [57], our method proves strong generalisation capacity in this experiment.

On the more challenging DomainNet dataset, firstly, we observe that in many tasks the adaptation actually harms the target performance (i.e., negative transfer). For instance, the images in "infograph" domain are largely occupied by texts, while the object to be recognised are highly abstracted or in various artistic styles. This could lead to failure in CNN learned features to capture the true commonalities in objects. The high accuracy of adaptation to the "real" domain can be mainly due to the use of ImageNet pre-trained model in the backbone networks. While for "quickdraw" domain, the perquisite from ImageNet model is of little help, and the network is forced to learn almost from scratch. Under such challenges, MIMTFL improves performance in 8 out of the 15 difficult tasks. While ADDA fails to bring any improvement on 7 tasks and MCD on 10, MIMTFL only fails on two, comparing to the source only model. Here since all the other methods focus on aligning the classifier output distributions using general CNN features, the effectiveness of our proposed MI-guided transferable feature learning is clearly proven.

Results of multiple-to-single tasks are in Table 3 and 4. While MIMTFL easily outperforms other methods on both datasets, we observe that it is especially outstanding for the difficult tasks. In the most difficult task APR→C in Office-Home experiments, MIMTFL+MEC improves the absolute target accuracy from source only model by 5.8%, bringing a 9.9% gain. In DomainNet experiments, our method creates new SOTA results in all of the six tasks, astoundingly boosting the source only model by 77%. Note that DCTN [52] and M^3SDA [38] are specifically designed multi-source adaptation algorithms that consider collabo-

Table 5. Results on Office-Home multi-to-multi UDA (all with ResNet-50 backbone).

Method	PR→AC	AP→CR	AR→CP	CP→AR	AC→PR	CR→AP	mean
Source only	59.7	68.9	69.0	72.8	74.9	76.0	70.2
SE [14]	54.6	62.4	63.4	70.1	70.0	73.4	65.6
DWT+MEC [43]	59.1	69.9	68.8	**76.2**	**78.8**	**77.6**	71.7
MIMTFL+MEC	**62.9**	**70.7**	**70.4**	75.7	76.1	77.3	**72.2**

Table 6. Results on DomainNet multi-to-multi UDA (all with ResNet-101 backbone).

Method	rs→iq	ir→qs	qr→is	is→qr	qs→ir	iq→rs	mean
Source only	14.0	17.6	31.0	35.4	36.2	42.0	29.4
ADDA [49]	4.2	2.9	15.9	20.6	27.7	17.2	14.7
MCD [44]	12.2	15.4	27.4	29.3	36.6	33.8	25.8
MIMTFL	**14.3**	**17.9**	**31.9**	**36.3**	**43.1**	**43.6**	**31.2**

rative learning among multiple source domains. This proves the superiority of our formulation of distributional uncertainty for multiple source domains.

Results of multi-to-multi tasks are shown in Table 5 and 6. In Office-Home experiments, our model is further proven to work well especially on the more difficult tasks, while other distribution alignment methods, such as SE, fail to even beat the source only model. In the more challenging DomainNet experiments, we observe negative transfer here again in ADDA and MCD. Since the scale of the target domain is multiplied, the adaptation becomes even harder. Results show that MIMTFL is able to improve the source model and performs the best in all the six tasks.

Domain Generalisation. We observe in results presented in Table 7 that our method easily outperforms existing state-of-the-art specialised DG methods. Specifically, we find that our shape learning is more effective than JiGen using jigsaw puzzle as transferable feature learning.

Table 7. Results on Office-Home multi-to-single DG (all with ResNet-18 backbone).

Method	ACP→R	ACR→P	APR→C	CPR→A	mean
D-SAMs[12]	71.5	69.2	44.4	**58.0**	60.8
JiGen[4]	72.8	71.5	47.5	53.0	61.2
MIMTFL	**74.4**	**73.1**	**51.1**	53.3	**63.0**

5 Conclusions

In this paper, we propose a theory grounded formulation for the definition of domain discrepancy using distributional uncertainty. We maximise the mutual information between source and target. In addition, we propose to enhance transferable feature learning in CNNs by balancing texture and non-texture feature learning with the measured uncertainty. To explicitly learn non-texture features, we propose a novel self-supervised object shape learning method, which can be used in parallel with many existing self-supervised visual feature learning methods. Our idea is thoroughly experimented and validated through extensive experiments.

References

1. Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D.: Mutual information neural estimation. In: ICML (2018)
2. Bhattacharyya, A., Fritz, M., Schiele, B.: Long-term on-board prediction of people in traffic scenes under uncertainty. In: CVPR (2018)
3. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (6), 679–698 (1986)
4. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: CVPR (2019)
5. Chen, C., Xie, W., Huang, W., Rong, Y., Ding, X., Huang, Y., Xu, T., Huang, J.: Progressive feature alignment for unsupervised domain adaptation. In: CVPR (2019)
6. Chen, X., Wang, S., Long, M., Wang, J.: Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In: ICML (2019)
7. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: CVPR (2019)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
9. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
10. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV (2015)
11. Dollár, P., Zitnick, C.L.: Fast edge detection using structured forests. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(8), 1558–1570 (2014)
12. D'Innocente, A., Caputo, B.: Domain generalization with domain-specific aggregation modules. In: German Conference on Pattern Recognition
13. Feng, D., Rosenbaum, L., Dietmayer, K.: Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection. In: IEEE Intelligent Transportation Systems Conference (ITSC) (2018)
14. French, G., Mackiewicz, M., Fisher, M.: Self-ensembling for visual domain adaptation. In: ICLR (2018)
15. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML (2015)
16. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research **17**(1), 2096–2030 (2016)
17. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: ICLR (2019)
18. Ghahramani, Z.: Probabilistic machine learning and artificial intelligence. Nature **521**(7553), 452–459 (2015)
19. Gölcü, D., Gilbert, C.D.: Perceptual learning of object shape. Journal of Neuroscience **29**(43), 13621–13629 (2009)
20. Goyal, P., Mahajan, D., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. In: ICCV (2019)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
22. Ilg, E., Cicek, O., Galesso, S., Klein, A., Makansi, O., Hutter, F., Brox, T.: Uncertainty estimates and multi-hypotheses networks for optical flow. In: ECCV (2018)

23. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016)
24. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: NIPS (2017)
25. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR (2018)
26. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22**(1), 79–86 (1951)
27. Larsson, G., Maire, M., Shakhnarovich, G.: Colorization as a proxy task for visual understanding. In: CVPR (2017)
28. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICML Workshop (2013)
29. Lee, H., Kim, H.E., Nam, H.: Srm: A style-based recalibration module for convolutional neural networks. In: ICCV (2019)
30. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML (2015)
31. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: NeurIPS (2018)
32. Long, M., Wang, J., Ding, G., Sun, J., Yu, P.S.: Transfer feature learning with joint distribution adaptation. In: ICCV (2013)
33. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: ICML (2017)
34. Nagy, Z., Braatz, R.D.: Distributional uncertainty analysis using power series and polynomial chaos expansions. Journal of Process Control **17**(3), 229–240 (2007)
35. Pan, Y., Yao, T., Li, Y., Wang, Y., Ngo, C.W., Mei, T.: Transferrable prototypical networks for unsupervised domain adaptation. In: CVPR (2019)
36. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS Workshop (2017)
37. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR (2016)
38. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: ICCV (2019)
39. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. In: ICLR Workshop (2017)
40. Qi, L., Wang, L., Huo, J., Zhou, L., Shi, Y., Gao, Y.: A novel unsupervised camera-aware domain adaptation framework for person re-identification. In: ICCV (2019)
41. Ringer, S., Williams, W., Ash, T., Francis, R., MacLeod, D.: Texture bias of cnns limits few-shot classification performance. arXiv preprint arXiv:1910.08519 (2019)
42. Robbins, H., Monro, S.: A stochastic approximation method. The annals of mathematical statistics **22**(3), 400–407 (1951)
43. Roy, S., Siarohin, A., Sangineto, E., Bulo, S.R., Sebe, N., Ricci, E.: Unsupervised domain adaptation using feature-whitening and consensus loss. In: CVPR (2019)
44. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR (2018)
45. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)
46. Straszecka, E.: Combining uncertainty and imprecision in models of medical diagnosis. Information Sciences **176**(20), 3026–3059 (2006)
47. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: ECCV (2016)

48. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NIPS (2017)
49. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR (2017)
50. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
51. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: CVPR (2017)
52. Xu, R., Chen, Z., Zuo, W., Yan, J., Lin, L.: Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In: CVPR (2018)
53. Zaech, J.N., Dai, D., Hahner, M., Van Gool, L.: Texture underfitting for domain adaptation. In: IEEE Intelligent Transportation Systems Conference (ITSC) (2019)
54. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018)
55. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
56. Zhao, H., Zhang, S., Wu, G., Moura, J.M., Costeira, J.P., Gordon, G.J.: Adversarial multiple source domain adaptation. In: NeurIPS (2018)
57. Zhou, K., Yang, Y., Hospedales, T.M., Xiang, T.: Deep domain-adversarial image generation for domain generalisation. In: AAAI (2020)
58. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107 (2002)