This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SHIFT15M: Fashion-specific dataset for set-to-set matching with several distribution shifts

Masanari Kimura
ZOZO Research
masanari.kimura@zozo.com

Takuma Nakamura
ZOZO Research
takuma.nakamura@zozo.com

Yuki Saito
ZOZO Research
yuki.saito@zozo.com

Figure 1. TSNE [30] visualization for the SHIFT15M.

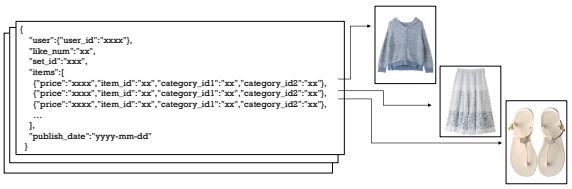

Figure 2. Overview of SHIFT15M dataset.

```python
from shift15m.datasets import NumLikesRegression

dataset = NumLikesRegression(root="./data", download=True)
(x_tr, y_tr), (x_te, y_te) = dataset.load_dataset(target_shift=True)
```

Figure 3. Minimum sample code using SHIFT15M data loader.

Abstract

Set-to-set matching is the problem of matching two different sets of items based on some criteria. Especially when each item in the set is high-dimensional, such as an image, set-to-set matching is treated as one of the applied problems to be solved by utilizing neural networks. Most machine learning-based set-to-set matching generally assumes that the training and test data follow the same distribution. However, such assumptions are often violated in real-world machine learning problems. In this paper, we propose SHIFT15M, a dataset that can be used to properly evaluate set-to-set matching models in situations where the distribution of data changes between training and testing. Some benchmark experiments show that the performance of naive methods drops due to the effects of the distribution shift. In addition, we provide software to handle SHIFT15M dataset in a very simple way: `https://github.com/st-tech/zozo-shift15m`*. The URL for the software will appear after this manuscript is published.*

1. Introduction

One of the key problems for fashion data analysis is set-to-set matching [1, 2, 23]. For example, we can consider a task that measures the degree of completion of an outfit by matching sets of clothing items (i.e., for two sets $A = \{\text{hat}, \text{shirt}, \text{skirt}\}$ and $B = \{\text{jacket}, \text{shoes}\}$, the matching score of A and B corresponds to the goodness of the outfit $A \cup B$). To solve this, we need to investigate neural networks that handle sets [13, 15, 26, 28, 35, 36, 39, 41].

Another common phenomenon in the domain of fashion is trend change. These phenomena are observed at various scales, ranging from annual trend changes such as fashionable colors to seasonal trend changes such as summer to winter clothing. In the field of machine learning, such an assumption can be defined as a distribution shift (or dataset shift) [11, 18, 19, 22, 24, 24, 27, 37].

Many machine learning problem settings assume that training and test data are independent and identically distributed (i.i.d). We assume that training examples $\{(\boldsymbol{x}_i^{tr}, y_i^{tr})\}_{i=1}^{n_{tr}}$ are independently and identically distributed (i.i.d.) according to some fixed but unknown distribution $p_{tr}(\boldsymbol{x}, y)$, which can be decomposed into the marginal distribution and the conditional probability distribution, i.e., $p_{tr}(\boldsymbol{x}, y) = p_{tr}(\boldsymbol{x})p_{tr}(y|\boldsymbol{x})$. We also denote the test examples by $\{(\boldsymbol{x}_i^{te}, y_i^{te})\}_{i=1}^{n_{te}}$ drawn from a test distribution $p_{te}(\boldsymbol{x}, y) = p_{te}(\boldsymbol{x})p_{te}(y|\boldsymbol{x})$.

Figure 4. Several sample images from SHIFT15M dataset.

Table 1. Statistics on the SHIFT15M dataset.

Property	Total	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
#sets	2,555,147	1,423	4,813	131,611	466,583	730,443	617,844	299,502	137,510	92,944	59,412	13,062
#items	15,218,721	8,327	29,140	756,532	2,644,564	4,305,802	3,731,864	1,853,647	855,036	576,022	373,549	84,238
mean set size	6.03	5.85	6.05	5.74	5.66	5.89	6.04	6.18	6.21	6.19	6.28	6.44
median set size	6.00	6.00	6.00	5.00	5.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00
mean #likes	26.98	0.94	2.00	15.74	16.84	23.24	37.37	35.67	32.41	24.89	21.34	16.01
median #likes	9.00	0.00	1.00	8.00	6.00	6.00	13.00	18.00	23.00	19.00	17.00	12.00
#unique users	193,574	289	571	16,922	52,283	80,290	49,441	18,854	7,511	4,442	2,739	853

Definition 1.1. (Covariate shift [25]) We consider that the two distributions $p_{tr}(\boldsymbol{x}, y)$ and $p_{te}(\boldsymbol{x}, y)$ satisfy the covariate shift assumption if the following conditions hold:

$$p_{tr}(\boldsymbol{x}) \neq p_{te}(\boldsymbol{x}), \quad p(y|\boldsymbol{x}) = p_{tr}(y|\boldsymbol{x}) = p_{te}(y|\boldsymbol{x}).$$

Definition 1.2. (Target shift [40]) We consider that the two distributions $p_{tr}(\boldsymbol{x}, y)$ and $p_{te}(\boldsymbol{x}, y)$ satisfy the target shift assumption if the following conditions hold:

$$p_{tr}(y) \neq p_{te}(y), \quad p(\boldsymbol{x}|y) = p_{tr}(\boldsymbol{x}|y) = p_{te}(\boldsymbol{x}|y).$$

To address these problems, we provide SHIFT15M, a real-world dataset that can handle the above two problem settings, that is, the set-to-set matching dataset with distribution shift. Our SHIFT15M dataset is built on data accumulated over the past 10 years in our fashion SNS. In this SNS, users could post combinations of their clothing items and other users could bookmark them as favorites. The data accumulated by this service, which has been in operation for a decade from 2010 to 2020, is very useful for dealing with distribution shifts in the fashion sector. Figure 2 shows an overview of the SHIFT15M dataset. Each column is a set of posted fashion items, with information such as the user who posted, the date of publication, and the price of each item. We hope that our SHIFT dataset will encourage research on set-to-set matching tasks under the distribution shift.

1.1. Contribution

Our contributions are summarized as follows:

- We propose SHIFT15M, a fashion-specific dataset that can properly evaluate models for set-to-set matching under the distribution shift assumptions. SHIFT15M also enables the performance evaluation of the model under various magnitudes of dataset shifts by switching the magnitude. Figure 4 shows several sample images from the SHIFT15M dataset.

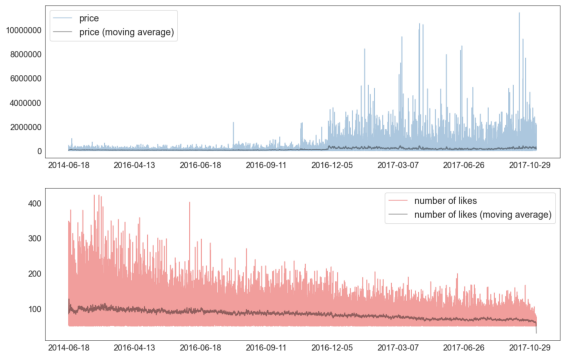

Figure 5. Top panel: trend of price for items included in SHIFT15M. Bottom panel: trend of number of likes for posted sets.

- We provide open-source software to handle the SHIFT15M dataset in a very simple way. Figure 3 shows the minimum sample code of our software;

- We propose first-step benchmark methods for set-to-set matching under distribution shift, numerical experiments show the usefulness of these methods.

2. Statistics on the SHIFT15M dataset

In this section, we present some statistics for our SHIFT15M dataset. First, Table 1 shows the overview of statistics on the SHIFT15M dataset. Since our fashion SNS was launched in 2010, the number of users and posts gradually increased from 2010, reaching a peak around 2014∼2015, and slowly decreasing until 2020, the year the service was terminated. Also, the number of items in a set tends to increase over the years, indicating that users tend to construct outfits with more and more items.

Top panel of Figure 5 shows the trend of price for items

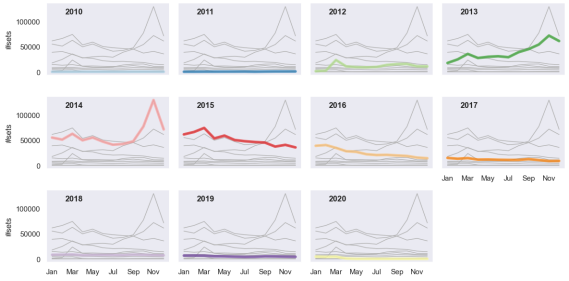

Figure 6. Trends of the number of posted sets by year

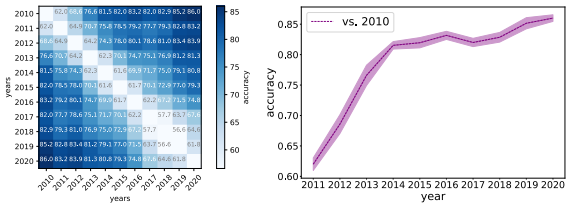

Figure 7. Covariate shift of image features.

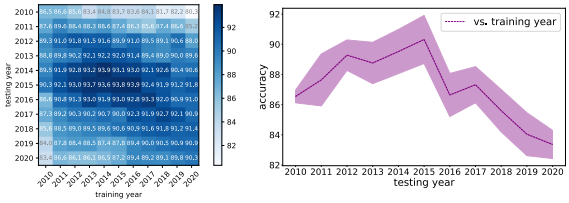

Figure 8. Category classification results under the covariate shift.

included in the SHIFT15M dataset. It can be seen that the fashion items posted by users are becoming more expensive every year. Bottom panel of Figure 5 shows the trend of number of likes for posted sets.

Figure 6 plots the trend of the number of posted sets by year. This figure shows that our fashion SNS, the source of the SHIFT15M dataset, was most active around 2014∼2015.

Finally, we confirm the covariate shift of the image features included in SHIFT15M. If covariate shift assumption 1.1 holds, we should be able to construct a classifier $f : \boldsymbol{x} \mapsto y = \{0, 1\}$, where $\boldsymbol{x}$ is the image feature of the item and y is the binary classification output for two years. Figure 7 shows the experimental results. The results show that classification between distant years (e.g., acc. of 2010 vs. 2020 is 0.85) is easier, while classification between close years (e.g., acc. of 2010 vs. 2011 is 0.62) is more difficult, indicating a gradual shift in image features. Figure 8 also shows the experimental results of item categorization when the training and test data were generated from different years. This figure shows that the closer the years of the training and test data are, the higher the classification accuracy.

3. Benchmarks

In this section, we introduce several numerical experiments on the SHIFT15M dataset.

3.1. Importance weighted set-to-set matching

As the benchmark strategy for the distribution shift adaptation on the set-to-set matching, we propose importance weighted set-to-set matching which is based on IWERM.

Definition 3.1. (Importance weighted ERM [25]) Importance Weighted Empirical Risk Minimization (IWERM) uses the density ratio $p_{te}(\boldsymbol{x})/p_{tr}(\boldsymbol{x})$ as the weighting function:

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \frac{p_{te}(\boldsymbol{x}_i^{tr})}{p_{tr}(\boldsymbol{x}_i^{tr})} \ell(h(\boldsymbol{x}_i^{tr}), y_i^{tr}). \quad (1)$$

Adopting the density ratio as the weighting function, as in Definition 3.1, leads to the following statistically important property.

Theorem 3.1. (Consistency of IWERM [25]) If we set $w(\boldsymbol{x}) = p_{te}(\boldsymbol{x})/p_{tr}(\boldsymbol{x})$ as the weighting function, the empirical error computed by the weighted ERM is consistent estimator of the expected error in the test distribution.

Using the above ideas, we propose a novel covariate shift adaptation method for set-to-set matching. Let $\mathcal{L}(\mathcal{V}, \mathcal{W}, f)$ be the K-pair-set loss [23] function for the set matching, which is defined as follows:

$$\mathcal{L}(\mathcal{V}, \mathcal{W}, f) = -\frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{K} \delta_{ij} \log \frac{\exp(f(\mathcal{V}_i, \mathcal{W}_j))}{\sum_{k=1}^{K} \exp(f(\mathcal{V}_i, \mathcal{W}_k))},$$

where δ is Kronecker's delta, and we can modify $\mathcal{L}(\mathcal{V}, \mathcal{W}, f)$ as follows:

$$\mathcal{L}_w(\mathcal{V}, \mathcal{W}, f) = -\frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{K} \delta_{ij} \Gamma_{i,j}^{p} \log \frac{\Gamma_{i,j}^{f}}{\sum_{k=1}^{K} \Gamma_{i,k}^{f}}, \quad (2)$$

where $\Gamma_{i,j}^{p} = e^{p(test|\mathcal{V}_i \cup \mathcal{W}_j)}$ and $\Gamma_{i,j}^{f} = e^{f(\mathcal{V}_i, \mathcal{W}_j)}$. This modification can be regarded as a weighting based on the probability that the pair is included in the test set. Here, we propose two weighting strategies:

$$\text{max-IW}: \; p(test|\mathcal{V}_i \cup \mathcal{W}_j) = \max_{\boldsymbol{x} \in \mathcal{V}_i \cup \mathcal{W}_j} w(\boldsymbol{x}),$$

$$\text{mean-IW}: \; p(test|\mathcal{V}_i \cup \mathcal{W}_j) = \frac{1}{|\mathcal{V}_i \cup \mathcal{W}_j|} \sum_{\boldsymbol{x} \in \mathcal{V}_i \cup \mathcal{W}_j} w(\boldsymbol{x}),$$

where $w(\boldsymbol{x})$ is the weighting function. Next, we approximate $w(\boldsymbol{x})$ by using unlabeled data from both p_{tr} and p_{te}. In IWERM, the squared error can be decomposed as follows:

$$\mathbb{E}_{p_{te}}\left[\Delta^2\right] = \mathbb{E}_{p_{tr}}\left[\hat{w}(\boldsymbol{x})\Delta^2\right] + \mathbb{E}_{p_{tr}}\left[(w(\boldsymbol{x}) - \hat{w}(\boldsymbol{x}))\Delta^2\right],$$

Table 2. Experimental results of the Fill-In-The-N-Blank with four candidates.

Models	2013	2014	2015	2016	2017
ERM [23]	0.924(±0.005)	0.907(±0.006)	0.886(±0.009)	0.865(±0.006)	0.855(±0.003)
ERM + mean-IW	0.924(±0.005)	0.917(±0.002)	0.886(±0.003)	0.866(±0.003)	0.860(±0.002)
ERM + max-IW	0.924(±0.005)	**0.921(±0.002)**	**0.896(±0.006)**	**0.871(±0.001)**	**0.865(±0.005)**

Table 3. Experimental results of the Fill-In-The-N-Blank with eight candidates.

Models	2013	2014	2015	2016	2017
ERM [23]	0.845(±0.000)	0.822(±0.001)	0.791(±0.005)	0.762(±0.008)	0.741(±0.004)
ERM + mean-IW	0.845(±0.000)	0.831(±0.008)	0.792(±0.002)	0.766(±0.004)	0.749(±0.002)
ERM + max-IW	0.845(±0.000)	**0.842(±0.004)**	**0.807(±0.003)**	**0.769(±0.005)**	**0.753(±0.005)**

where $\Delta^2 = \|f(\boldsymbol{x}) - y\|^2$ and $\hat{w}(\boldsymbol{x})$ is the approximator of the weighting function $w(\boldsymbol{x})$. Second term is bounded as

$$\mathbb{E}_{p_{tr}}\Big[(w(\boldsymbol{x}) - \hat{w}(\boldsymbol{x}))\Delta^2\Big] \leq \frac{1}{2}\Big(\mathbb{E}_{p_{tr}}\Big[\Delta^2\Big] + \mathbb{E}_{p_{tr}}\Big[(w(\boldsymbol{x}) - \hat{w}(\boldsymbol{x}))^2\Big]\Big). \quad (3)$$

Let s is the indicator of the distributions, where $s = 1$ corresponds to the train distribution and $s = 0$ corresponds to the test distribution, and we assume that $p(s) = 0.5$. Then, we also assume that

$$p(\boldsymbol{x}|s) = \begin{cases} p_{tr}(\boldsymbol{x}) & (s = 1), \\ p_{te}(\boldsymbol{x}) & (s = 0). \end{cases} \quad (4)$$

Then, we have $w(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|s=0)}{p(\boldsymbol{x}|s=1)}$. Let $g(\boldsymbol{x})$ be the optimal source discriminator which identifies whether $\boldsymbol{x}$ is generated p_{tr} or p_{te}. Then, we can write as $g(\boldsymbol{x}) = p(s = 1|\boldsymbol{x}) = \frac{1}{1+w(\boldsymbol{x})}$. Suppose that the density ratio $p_{te}(\boldsymbol{x})/p_{tr}(\boldsymbol{x})$ is bounded by $\beta > 0$, we have $\frac{1}{1+\beta} \leq g(\boldsymbol{x}) \leq 1$ for all $\boldsymbol{x}$. From the unlabeled data generated from p_{tr} and p_{te}, we can learn the estimator $\hat{g}$ of g. Then, we can write the weight estimation term as

$$\begin{aligned}
&\mathbb{E}_{p_{tr}}\Big[(w(\boldsymbol{x}) - \hat{w}(\boldsymbol{x}))^2\Big] = \mathbb{E}_{p_{tr}}\left[\left(\frac{g(\boldsymbol{x}) - \hat{g}(\boldsymbol{x})}{g(\boldsymbol{x})\hat{g}(\boldsymbol{x})}\right)^2\right] \\
&\leq (1+\beta)^4 \mathbb{E}_{p_{tr}}\Big[(g(\boldsymbol{x}) - \hat{g}(\boldsymbol{x}))^2\Big] \\
&= (1+\beta)^4 \mathbb{E}_{p_{te}}\left[(g(\boldsymbol{x}) - \hat{g}(\boldsymbol{x}))^2 \frac{p_{tr}(\boldsymbol{x})}{p_{te}(\boldsymbol{x})}\right] \\
&\leq 2(1+\beta)^4 \mathbb{E}_{p_{te}}\Big[(g(\boldsymbol{x}) - \hat{g}(\boldsymbol{x}))^2\Big] \\
&= 2(1+\beta)^4 \left\{\mathbb{E}_{p_{te}}\Big[(s - g(\boldsymbol{x}))^2\Big] - \mathbb{E}_{p_{te}}\Big[(g(\boldsymbol{x}) - \hat{g}(\boldsymbol{x}))\Big]\right\}.
\end{aligned}$$

This indicates that the weighting function is approximated by the function $g(\boldsymbol{x})$.

3.2. Results of numerical experiments

We introduce benchmark results for a set-to-set matching under the covariate shift. The model architecture is same as the previous work [23], which is based on the architecture of Transformer [15, 20, 33]. Our task can be considered an extended version of a standard task, Fill-In-The-Blank [8], which requires us to select an item that best extends an outfit from among four candidates. Because selecting a set corresponds to filling multiple blanks, we consider the set matching problem as Fill-In-The-N-Blank [23]. To construct the correct pair of sets to be matched, we randomly halve the given outfit $\mathcal{O}$ into two non-empty proper subsets $\mathcal{V}$ and $\mathcal{W}$, as follows: $\mathcal{O} \to \{\mathcal{V}, \mathcal{W}\}$, where $\mathcal{V} \cap \mathcal{W} = \emptyset$.

Tables 2 and 3 show the experimental results of the Fill-In-The-N-Blank with four and eight candidates. In these experiments, data from 2013 are used as training data and data from 2013∼2017 are used as test data. ERM refers to empirical risk minimization [6, 31, 32], which assumes that $p_{tr}(\boldsymbol{x}) = p_{te}(\boldsymbol{x})$. From these results, we can see that the covariate shift adaptive set-to-set matching methods can achieve better performances than the ordinal ERM.

4. Related works and conclusion

Several distribution shift datasets exist for general classification and regression tasks where the input is a vector. WILDS [14] is the collection of benchmark datasets [3–5, 7, 9, 12, 17, 21, 29, 38] under the distribution shift, including histopathological images, satellite images or sequence of source code tokens. PACS [16] and Office-Home [34] adopt the image style to differentiate distributions, and VLCS [10] takes data collected independently from four sources as environments. Also, DomainNet [42] extends PACS to a far larger scale, consisting of more domains and categories.

We believe that our SHIFT15M is a very useful dataset for evaluating the still underdeveloped task of set-to-set matching under natural distribution shifts.

References

[1] Ognjen Arandjelović. Discriminative extended canonical correlation analysis for pattern set matching. *Machine Learning*, 94:353–370, 2014. 1

[2] Yunsheng Bai, Hao Ding, Yizhou Sun, and Wei Wang. Convolutional set matching for graph similarity. *arXiv preprint arXiv:1810.10866*, 2018. 1

[3] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018. 4

[4] Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020. 4

[5] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, 2019. 4

[6] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer School on Machine Learning*, pages 169–207. Springer, 2003. 4

[7] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4

[8] Guillem Cucurull, Perouz Taslakian, and David Vazquez. Context-aware visual compatibility prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12617–12626, 2019. 4

[9] Etienne David, Simon Madec, Pouria Sadeghi-Tehran, Helge Aasen, Bangyou Zheng, Shouyang Liu, Norbert Kirchgessner, Goro Ishikawa, Koichi Nagasawa, Minhajul A Badhon, Curtis Pozniak, Benoit de Solan, Andreas Hund, Scott C. Chapman, Frederic Baret, Ian Stavness, and Wei Guo. Global wheat head detection (gwhd) dataset: a large and diverse dataset of high-resolution rgb-labelled images to develop and benchmark wheat head detection methods. *Plant Phenomics*, 2020, 2020. 4

[10] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013. 4

[11] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 1

[12] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 4

[13] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 1

[14] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. 4

[15] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019. 1, 4

[16] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 4

[17] Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664*, 2021. 4

[18] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021. 1

[19] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012. 1

[20] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 4

[21] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. 4

[22] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008. 1

[23] Yuki Saito, Takuma Nakamura, Hirotaka Hachiya, and Kenji Fukumizu. Exchangeable deep neural networks for set-to-set matching and learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII*, pages 626–646. Springer, 2020. 1, 3, 4

[24] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution

generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021. 1

[25] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000. 2, 3

[26] Maximilian Soelch, Adnan Akhundov, Patrick van der Smagt, and Justin Bayer. On deep set learning and the choice of aggregations. In *Artificial Neural Networks and Machine Learning–ICANN 2019: Theoretical Neural Computation: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part I 28*, pages 444–457. Springer, 2019. 1

[27] Adarsh Subbaswamy, Roy Adams, and Suchi Saria. Evaluating model robustness and stability to dataset shift. In *International Conference on Artificial Intelligence and Statistics*, pages 2611–2619. PMLR, 2021. 1

[28] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022. 1

[29] J. Taylor, B. Earnshaw, B. Mabey, M. Victors, and J. Yosinski. Rxrx1: An image set for cellular morphological variation across many experimental batches. In *International Conference on Learning Representations (ICLR)*, 2019. 4

[30] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1

[31] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013. 4

[32] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999. 4

[33] 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. 4

[34] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 4

[35] Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A Osborne. On the limitations of representing functions on sets. In *International Conference on Machine Learning*, pages 6487–6494. PMLR, 2019. 1

[36] Edward Wagstaff, Fabian B Fuchs, Martin Engelcke, Michael A Osborne, and Ingmar Posner. Universal approximation of functions on sets. *Journal of Machine Learning Research*, 23(151):1–56, 2022. 1

[37] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 1

[38] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications*, 2020. 4

[39] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017. 1

[40] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013. 2

[41] Yan Zhang, Jonathon Hare, and Adam Prugel-Bennett. Deep set prediction networks. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[42] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 4