Transfer Learning for Fine-grained Classification Using Semi-supervised Learning and Visual Transformers

Manuel Lagunas*
mlgns@amazon.com

Brayan Impata*
biimpata@amazon.com

Victor Martinez
vicmg@amazon.com

Virginia Fernandez
virfer@amazon.com

Christos Georgakis
georgak@amazon.com

Sofia Braun
brasofia@amazon.com

Felipe Bertrand
felipb@amazon.com

Abstract

Fine-grained classification is a challenging task that involves identifying subtle differences between objects within the same category. This task is particularly challenging in scenarios where data is scarce. Visual transformers (ViT) have recently emerged as a powerful tool for image classification, due to their ability to learn highly expressive representations of visual data using self-attention mechanisms. In this work, we explore Semi-ViT, a ViT model fine tuned using semi-supervised learning techniques, suitable for situations where we have lack of annotated data. This is particularly common in e-commerce, where images are readily available but labels are noisy, nonexistent, or expensive to obtain. Our results demonstrate that Semi-ViT outperforms traditional convolutional neural networks (CNN) and ViTs, even when fine-tuned with limited annotated data. These findings indicate that Semi-ViTs hold significant promise for applications that require precise and fine-grained classification of visual data.

1. Introduction

In recent years, the development of deep neural networks has led to significant advancements in the field of computer vision [16]. One such architecture is the Visual Transformer (ViT) [5], which utilizes the self-attention mechanism to model long-range dependencies between image features. Unlike traditional convolutional neural networks (CNN) [7, 10, 26], which rely on handcrafted hierarchical feature extraction, visual transformers can learn global spatial relationships among image features in a more efficient and effective manner. This has enabled them to outperform state-of-the-art methods on various visual recognition tasks [18]. However, in real-world scenarios labeled data can be scarce and expensive to obtain. Therefore, semi-

* Joint first authors.

supervised learning (SSL) [40] has emerged as a powerful technique for leveraging unlabeled data to improve the performance of deep neural networks. CNN methods have significantly advanced the field [1, 3, 15, 27, 32] while ViT architectures have only recently demonstrated promising results [2, 33] with SSL.

In this paper, we investigate the effectiveness of SSL when used with ViT architectures. Specifically, we utilize the Semi-ViT architecture [2] to conduct transfer learning for fine-grained classification of e-commerce data. The use of e-commerce data presents a unique advantage for SSL as unlabeled images are readily available. However, the labels associated are often noisy or absent altogether. Traditionally, this issue has been addressed through the use of manual curators, which can be costly and predominantly accessible for established marketplaces. In emerging markets, such as in Latin America, the scarcity of reliable labelled data poses an even greater challenge.

We collect three datasets from e-commerce data containing labeled and unlabeled images. We perform fine-grained classification on the neck style of a vest (*Vest Neck Style*), the pattern of a phone case (*Phone Case Pattern*), and the pattern of aprons and food bibs (*Apron Food Bib Pattern*). Each of the datasets contains 29K, 30K, and 33K labeled images, and 227K, 287K, 284K unlabeled images, respectively. Labels were gathered using crowd-sourced methods. We fine tune three different models, the well-known ResNet architecture [10], a ViT, and a Semi-ViT architecture; all of them pretrained on ImageNet [4]. For the ViT and Semi-ViT architectures, we additionally set different labeled data regimes where they are additionally fine-tuned using 25%, 50%, and 75% of the labeled data for each of the datasets. In total, we train 9 different models for each task.

2. Related Work

Visual Transformers   Visual Transformers (ViT) have recently achieved state-of-the-art performance in many computer vision tasks [5, 19, 31]. They adapt self-attention

mechanisms to the image domain, allowing to better capture long-range dependencies and contextual information. ViTs have been extended with knowledge distillation [29], using token-level and patch-level transformer layers [8], or progressively downsampling the image [37]. A comprehensive review on ViTs can be found in the work of Khan et al. [11]. In this work, we explore and compare the performance of ViT architectures and traditional CNNs, fine-tuned with supervised and semi-supervised techniques for fine-grained classification.

Transfer Learning Transfer learning leverages pre-trained models and adapts them to new domains [13, 22, 25, 41]. Yosinski et al. [35, 36] investigate the transferability of features learned by deep neural networks on different tasks, demonstrating their effectiveness. Transfer learning has also been applied for object detection and semantic segmentation [6], and in unsupervised domain adaptation [30]. We use fine tuning methods in ResNet, ViT, and Semi-Vit [2] architectures to perform fine-grained classification in three different datasets.

Semi-Supervised Learning Semi-Supervised Learning (SSL) uses labeled and unlabeled data to improve model performance when labeled data is scarce [14, 34, 38]. It leverages intelligent data augmentation techniques paired with consistency regularization to improve performance [1, 20, 32, 39]. Other approaches rely on pseudo-labelling [17], teacher-student models [24, 28], ensembles [15], or adversarial training [12, 21]. We apply SSL to e-commerce images, where labeled data is expensive, but we can retrieve a large amount of unlabeled samples. We demonstrate the applicability of SSL and showcase its efficacy at reducing the need for labeled data.

3. Data Collection

Our goal is to compare the performance of traditional CNNs and ViT architectures on e-commerce images that allow us to have large amounts of unlabelled data. However, labelled data is noisy or nonexistent, therefore we would need crowd-sourcing tasks to label them.

We use three different datasets to perform fine-grained classification: predict the neck style of vests (*Vest Neck Style*), the pattern of phone cases (*Phone Case Pattern*), and the pattern of aprons and food bibs (*Apron Food Bib Pattern*). All images come from Amazon's marketplace. To label images we rely on Amazon Mechanical Turk. Each labelled image has answers from three different people. We consider an image as labelled if two or more labels are the same, otherwise, we would discard the labels and consider it unlabelled. To ensure the quality of the annotations, we only allowed workers that previously had passed a prelim-

Dataset	Labelled	Unlabelled	Classes
VestNeckStyle	29K	227K	13
PhoneCasePattern	37K	287K	21
ApronFoodBibPattern	39K	284K	26

Table 1. Summary of the number of images that are labelled, unlabelled, and the number of classes for each of the datasets that we collected from e-commerce data. All the datasets have a class named as *other*. This class is used to label images that do not belong to the aforementioned data i.e., in *Vest Neck Style*, an image showing something that is not a vest.

Figure 1. Examples of the images collected for our datasets. From top to bottom, we can see examples for *Vest Neck Style*, *Phone Case Pattern*, and *Apron Food Bib Pattern*.

inary test task. Table 1 shows a summary on the dataset statistics. An example of images we collected for each of the three tasks can be seen in Figure 1.

4. Methodology

Our goal is to compare the performance of CNNs and modern ViT architectures. In addition, we will also study the influence of semi-supervised learning with ViT (Semi-ViT) during training using all available labeled data, and investigate the influence of more restrictive data regimes in ViT's and Semi-ViT's performance.

Models We employ three different models: A CNN (a ResNet [10]) model, a ViT [5] model, and a semi-ViT model [2] (a ViT model trained also with SSL). For the

CNN we use the well-know ResNet18 [10] model. This model has shown remarkable performance across different computer vision tasks while being the less parameter-heavy ResNet model. In the ViT architecture, we have relied on Masked Autoencoders (MAE) [9] ViT-Base model. Following a similar approach to ResNet, this is the most parameter efficient MAE. For the Semi-ViT model [2], we also use the same MAE ViT-Base model. During the semi-supervised stage, an exponential moving average (EMA)-Teacher framework is adopted together with a probabilistic pseudo mixup [39] method that allows for better regularization by interpolating unlabeled samples and their pseudo labels.

Data The labeled data is separated into *train*, *validation*, and *test*. Since we gather e-commerce data, the underlying label distribution is unknown, the *train* and *validation* sets do not have a uniform distribution of the labels across images. We sample the *test* set to have the same number of images for all labels. The distribution of the data is roughly 75%, 15%, and 10% for *train*, *validation*, and *test* respectively.

Fine tuning All three models have been pre-trained on ImageNet [4]. We fine-tune every model on 100% of the labeled samples. For ViT and Semi-ViT, we also experiment tuning them with 25%, 50%, and 75% of the training data. The validation and test sets remains the same to have a fair comparison of performance. For Semi-ViT, we fine tune and later perform semi-supervised learning using both labeled and unlabelled data. We orchestrate the fine tuning of all models using *SageMaker g5* instances, we rely on their hyper-parameter tuner with Bayesian search [23] configured to search hyper-parameters around the original values given in the training code of the three models. On Average, training a model took 1h for ResNet, 4h for ViT, and 12h for Semi-ViT.

5. Results

In this Section we present the results obtained by the fine-tuned models in all three tasks. Table 2 shows results for all models in the *Vest Neck Style*, *Phone Case Pattern*, and *Apron Food Bib Pattern* task. For each, we show the percentage of data used for training, the accuracy top-1 (Acc@1) and top-5 (Acc@5), and the cross entropy (CE) error. All values are obtained in the test set.

We observe that the Semi-ViT model outperforms others in all three tasks, showcasing the benefits of SSL to achieve better generalization. On the other hand, ResNet18 is the model that obtains worst performances. Other interesting finding is that, in general, the performance of ViT/50% is on par with Semi-ViT/25%. Thus, using half the amount of an-

VEST NECK STYLE			
Method	Acc@1	Acc@5	Loss (CE)
ResNet18/100%	79.736	98.775	0.615
ViT/100%	81.244	99.246	0.631
ViT/75%	81.056	98.586	0.661
ViT/50%	80.584	98.586	0.704
ViT/25%	76.343	96.418	0.811
Semi-ViT/100%	**85.297**	**99.246**	**0.549**
Semi-ViT/75%	83.129	98.963	0.573
Semi-ViT/50%	81.433	98.115	0.637
Semi-ViT/25%	81.244	97.455	0.679

PHONE CASE PATTERN			
Method	Acc@1	Acc@5	Loss (CE)
ResNet18/100%	72.931	97.034	0.923
ViT/100%	79.005	98.613	0.685
ViT/75%	77.810	98.135	0.732
ViT/50%	77.571	97.513	0.758
ViT/25%	74.175	97.131	0.871
Semi-ViT/100%	**81.540**	**98.613**	**0.618**
Semi-ViT/75%	80.010	98.374	0.664
Semi-ViT/50%	79.149	98.422	0.700
Semi-ViT/25%	76.279	97.465	0.805

APRON FOOD BIB PATTERN			
Method	Acc@1	Acc@5	Loss (CE)
ResNet18/100%	73.766	95.598	0.987
ViT/100%	78.079	97.688	0.739
ViT/75%	78.301	97.821	0.757
ViT/50%	75.056	96.754	0.869
ViT/25%	69.898	94.309	1.070
Semi-ViT/100%	**81.814**	97.732	**0.663**
Semi-ViT/75%	81.547	**97.866**	0.685
Semi-ViT/50%	77.234	97.110	0.815
Semi-ViT/25%	73.499	94.620	0.981

Table 2. Results using the ResNet18, ViT, and Semi-ViT models with different data regimes on the three datasets obtained from e-commerce sources (images) and crowd-sourced experiments (labels). We can see how Semi-ViT, the visual transformer model that also uses SSL techniques, obtains the best performance for all tasks (green highlighted). Moreover, we can see that the Semi-ViT model with scarce data regimes (e.g., 50% of the training data), which is a common scenario in e-commerce, obtain performances comparable to ViT models trained with double the amount of data.

notated data. Similarly, the performance of Semi-ViT/50%, is on par or superior to ViT/100%.

5.1. Performance per class

If we analyze the performance per class for each of the three tasks using Semi-ViT/100%, we observe that the models, in general, achieve high values of accuracy top-1, usually above 90% per label. We find performances below 50% on the labels that are underrepresented in the training set. In the case of the *Apron Food Bib Pattern* task, three worse performing labels: *paisley*, *patched*, and *trellis*; show an accuracy top-1 of 18.18%, 11.11%, 40.00% respectively. However, they only represent 0.17%, 0.07%, and 0.56% of the training data. In the case of *Phone Case Pattern*, we also observe *paisley* as an underperforming class, nevertheless, in training it only represented 0.71% of the data. One interesting case, is the *inconclusive* label, representing phone cases that could not be labeled in any other class. The model shows a 45.78% accuracy top-1 even when the class represents 3.17% of the samples. We argue that this low performance comes due to the fact that this is a "wildcard" class in which we may find high degrees of variability between images. Last, for *Vest Neck Style*, we only find the *band collar* label to have low performance (17.65%), this label only represented 0.82% of the training data.

5.2. Performance by marketplace

When downloading e-commerce data, we also stored metadata such as the marketplace where this image was located. In Latin America, we obtained data from Mexico and Brazil. For *Vest Neck Style*, we found the accuracy top-1 of the Semi-ViT/100% model to be on par to other regions, with 85.18% in Mexico, and 82.92% in Brazil. For *Phone Case Pattern*, the accuracy top-1 is 80.56% in Mexico, and 60.00% in Brazil. In this case, we found significantly less images in Brazil than in others. Therefore, we have few images to train, resulting in the model not properly learning the characteristics of the images for this marketplace. For *Apron Food Bib Pattern* we find accuracy top-1 to be 78.88% in Mexico, and 81.91% in Brazil; on par with other marketplaces.

5.3. Influence on the amount of unlabeled data

Since we use e-commerce data, obtaining new unlabeled images is a fairly easy process. Therefore, we investigate the effect of having additional unlabeled data in models' performance. We selected the Semi-ViT model using 50% of the labeled data (Semi-ViT/50%), and train it using SSL with 200%, 300%, and 400% of the unlabeled data. In the end, we have additionally collected a total of 1139K unlabeled images for the *Apron Food Bib Pattern* task.

We observe in Table 3 that adding extra unlabeled images yields results that improve performance in the case of 200% and 300% of the available unlabeled images. However, it seems to collapse for 400% where performance decreases and is on par with using 100% of the unlabeled data.

ADDITIONAL UNLABELED			
Unlabeled	Acc@1	Acc@5	Loss (CE)
284K (100%)	77.234	**97.110**	0.815
569K (200%).	78.301	**97.110**	0.786
854K (300%)	**79.146**	96.843	**0.781**
1139K (400%)	77.679	96.843	0.801

Table 3. Results for *Apron Food Bib Pattern* using the Semi-ViT/50% model and increasing the amount of unlabeled data. We can see how the performance in terms of accuracy and loss improves while adding more data. However, it collapses with 400% of unlabeled data. We hypothesize that this may be due to the percentage of labeled data being significantly smaller with respect to unlabeled data, thus, unlabeled data driving the training and not allowing to further generalize.

We argue that the labeled data represents a very small percentage compared to unlabeled data, thus, the latter data drives model's training yielding worse performance.

6. Conclusion

We have run experiments with three models: ResNet, ViT, and Semi-Vit on three datasets that were collected from e-commerce data. The use of e-commerce data provided a realistic setting to assess the impact of SSL with a combination of labeled and unlabeled images. Our experiments showed that Semi-ViT can effectively leverage the benefits of SSL to improve the performance in fine-grained classification tasks compared to other architectures. We can see how Semi-ViT obtains accuracy values similar to ViT models that have trained with double the amount of data. Those results are shared across datasets showing promise to reduce the need for labeled data while maintaining high performance. Nevertheless, there are open venues worth exploring: One could aim to integrate the noise from the crowd-sourced labels into the loss function to create more robust models, or explore the creation of SSL techniques for multi-attribute models in which all three tasks would be performed by a single model.

References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 1, 2

[2] Zhaowei Cai, Avinash Ravichandran, Paolo Favaro, Manchen Wang, Davide Modolo, Rahul Bhotika, Zhuowen Tu, and Stefano Soatto. Semi-supervised vision transformers at scale. *arXiv preprint arXiv:2208.05688*, 2022. 1, 2, 3

[3] Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Zhuowen Tu, and Stefano Soatto. Exponential moving average normalization for self-supervised and semi-

supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 194–203, 2021. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 3

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[6] 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

[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 1

[8] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021. 2

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3

[10] 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, 2, 3

[11] 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. 2

[12] Abhishek Kumar, Prasanna Sattigeri, and Tom Fletcher. Semi-supervised learning with gans: Manifold invariance with improved inference. *Advances in neural information processing systems*, 30, 2017. 2

[13] Manuel Lagunas and Elena Garces. Transfer learning for illustration classification. *arXiv preprint arXiv:1806.02682*, 2018. 2

[14] Manuel Lagunas, Sandra Malpica, Ana Serrano, Elena Garces, Diego Gutierrez, and Belen Masia. A similarity measure for material appearance. *ACM Transactions on Graphics (TOG, Proc. SIGGRAPH)*, 2019. 2

[15] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 1, 2

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[17] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 2

[18] Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. A survey of visual transformers. *arXiv preprint arXiv:2111.06091*, 2021. 1

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[20] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 2

[21] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 2

[22] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014. 2

[23] Valerio Perrone, Huibin Shen, Aida Zolic, Iaroslav Shcherbatyi, Amr Ahmed, Tanya Bansal, Michele Donini, Fela Winkelmolen, Rodolphe Jenatton, Jean Baptiste Faddoul, et al. Amazon sagemaker automatic model tuning: Scalable gradient-free optimization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3463–3471, 2021. 3

[24] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, pages 135–152, 2018. 2

[25] Matthia Sabatelli, Mike Kestemont, Walter Daelemans, and Pierre Geurts. Deep transfer learning for art classification problems. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[27] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 1

[28] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2

[29] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2

[30] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceed-*

ings of the IEEE conference on computer vision and pattern recognition, pages 7167–7176, 2017. 2

[31] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12894–12904, 2021. 1

[32] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020. 1, 2

[33] Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Svformer: Semi-supervised video transformer for action recognition. *arXiv preprint arXiv:2211.13222*, 2022. 1

[34] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 2

[35] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 2

[36] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015. 2

[37] Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip HS Torr, Wayne Zhang, and Dahua Lin. Vision transformer with progressive sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 387–396, 2021. 2

[38] Brayan S Zapata-Impata and Pablo Gil. Prediction of tactile perception from vision on deformable objects. 2020. 2

[39] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2, 3

[40] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005. 1

[41] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 2