Towards Generalizing to Unseen Domains with Few Labels

Supplementary Material

6. Datasets Description

PACS: PACS dataset consists of four domains: photo (1,670 images), art painting (2,048 images), cartoon (2,344 images) and sketch (3,929 images), and seven classes: dog, elephant, giraffe, guitar, horse, house, and person.

OfficeHome: The OfficeHome dataset also consists of four domains: art, clipart, product, and the real world with 65 different classes whose total number of images adds up to around 15,500.

Digits-DG: The Digits-DG dataset consists of four domains: MNIST, MNIST-M, SVHN, and SYN, and 10 classes from 0 to 9. MNIST comprises hand-written digits, MNIST-M is a variant of MNIST with a blended background, SVHN has street-view house numbers and SYN consists of synthetic digits.

VLCS: VLCS dataset consists of four domains: Caltech, Labelme, Pascal, and Sun with 5 classes: bird, car, chair, dog, and person

Terra Incognita: This dataset contains images belonging to four domains collected from different locations namely locations 38, 43,46 and 100 with 10 classes: bird, bobcat, cat, coyote, dog, empty, opossum, rabbit, raccoon, and squirrel.

Figure 6 displays randomly selected samples from each dataset.

7. Comparison with DG baselines

We choose ERM [37], MixUp [54], GroupDRO [29], CrossGrad [31], DAELDG [60] , DDAIG [58] and DomainMix [34] as DG baselines to compare our method on PACS, OfficeHome, Digits-DG, VLCS and TerraIncognita datasets under both 5 labels (see Tab. 5) and 10 labels (see Tab. 6) settings.

8. Additional class similarity matrices

To show the effectiveness of our algorithm, we show cosine similarity between the class means on PACS dataset in the main manuscript. In this supplementary, we also show this on Digits-DG dataset (see Fig. 7). We see that, our losses encourage orthogonality among features with different class labels, and hence better discrimination in the feature space under different domain shifts and limited labels. Especially in Mnist-m, and Syn domains the similarities are notably higher compared to the Fixmatch baseline.

9. Confusion Matrices

We plot the confusion matrices comparison between the Fixmatch baseline and our method on PACS (Fig. 8), and Digits-DG (Fig. 9) datasets. Compared to FixMatch, our approach shows improved class-wise accuracy in both datasets.

10. Additional Feature Visualizations

In addition to the feature visualization on PACS dataset (main manuscript), we visualize features on Digits-DG dataset in Fig. 10 for Fixmatch and our method. In our approach, we observe that intra-class features are closer and inter-class features are far apart. We note that, compared to the baseline, the classes are well-separated in our method. Also, the comparison of cosine similarity of mean class features in Fig. 7 validates this further.

11. Comparison with other backbones

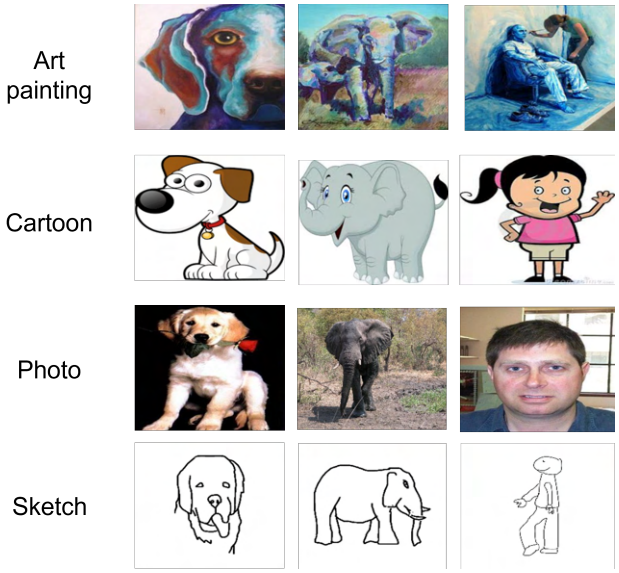
For a fair comparison, we used the same backbone (ResNet18) as used in StyleMatch. To evaluate our method further, we analyze the performance OfficeHome dataset with several stronger backbones such as ImageNet pre-trained ResNet50, ResNet101, ViT-S/32, ViT-B/32, and CLIP pre-trained ViT-B/32 (CLIP-B/32) in Tab. 7. Our method *consistently outperforms* the baseline even with other stronger backbones.

12. Scaling of performance with # labels

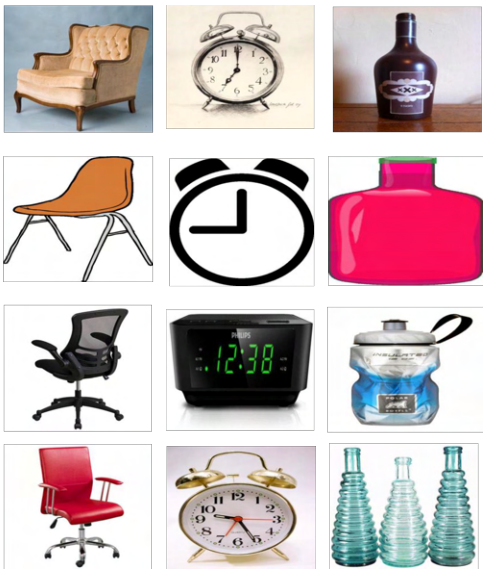
Tab. 8 provides results with increasing number of per-class labels. We see that the performance of our method improves upon increasing the number of per-class labels and in all per-class labels settings, it is higher than the other methods. Furthermore, with just 100 per-class labels, our method performance is better than the fully supervised ERM (i.e. with all labels) that obtains $80.0_{\pm 0.5}$.

13. Runtime and Memory overhead comparison

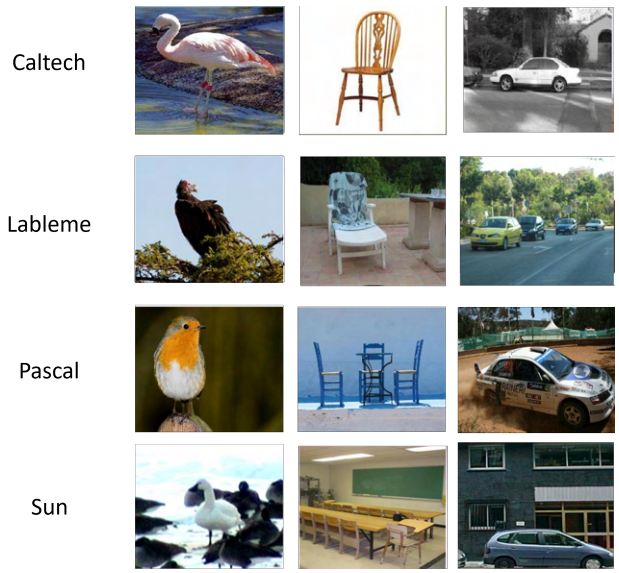
To evaluate the effectiveness of our method, we compare the runtime and memory overhead (Tab. 9.) Our method adds a little runtime overhead (20%) over the baseline FixMatch compared to the existing SSDG method, StyleMatch (117%) which has the same baseline (FixMatch).



(a) PACS

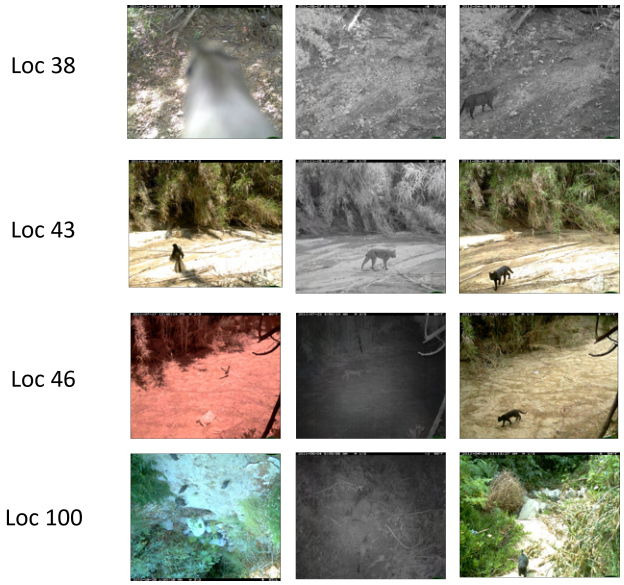


(b) OfficeHome



(c) VLCS

(d) Digits-DG



(e) Terra Incognita

Figure 6. Example images from different DG datasets used in our experiments.

Model	PACS	OH	VLCS	DigitsDG	TerraInc.
ERM	51.2 ± 3.0	51.7 ± 0.6	67.2 ± 1.8	22.7 ± 1.0	22.9 ± 3.0
MixUp	45.3 ± 3.8	52.7 ± 0.6	69.9 ± 1.3	21.7 ± 1.9	21.0 ± 2.9
GroupDRO	48.2 ± 3.6	53.8 ± 0.6	69.8 ± 1.2	23.1 ± 1 9	22.4 ± 3.1
CrossGrad	50.6 ± 3.4	51.6 ± 0.9	68.1 ± 1.6	22.8 ± 0.4	21.4 ± 2.3
DAELDG	42.7 ± 2.7	47.3 ± 0.6	61.7 ± 1.9	22.3 ± 1.0	25.0 ± 3.0
DDAIG	50.5 ± 3.0	50.6 ± 0.7	65.2 ± 2.2	23.2 ± 1.8	31.7 ± 3.1
DomainMix	46.3 ± 3.5	49.9 ± 0.6	68.3 ± 0.7	20.6 ± 1.4	22.9 ± 0.9

Table 5. DG accuracy (%) under SSDG settings (5 labels per class). Average over 5 independent seeds is reported.

Model	PACS	OH	VLCS	DigitsDG	TerraInc.
ERM	59.8 ± 2.3	56.7 ± 0.8	68.0 ± 0.3	29.1 ± 2.9	23.5 ± 1.2
MixUp	58.5 ± 2.2	57.2 ± 0.6	69.6 ± 1.0	29.7 ± 3.1	24.8 ± 3.3
GroupDRO	57.3 ± 1.2	57.8 ± 0.4	69.4 ± 0.9	31.5 ± 2.5	25.8 ± 3.3
CrossGrad	59.7 ± 1.5	56.7 ± 0.4	67.9 ± 0.6	30.3 ± 2.7	22.6 ± 0.9
DAELDG	53.7 ± 2.1	54.8 ± 0.3	68.3 ± 1.3	28.6 ± 1.5	25.5 ± 2.6
DDAIG	59.6 ± 1.6	55.1 ± 0.1	68.5 ± 1.0	29.4 ± 3.0	23.5 ± 3.0
DomainMix	58.0 ± 1.9	55.4 ± 0.4	69.5 ± 0.8	24.6 ± 1.1	23.3 ± 2.1

Table 6. DG accuracy (%) under SSDG settings (10 labels per class). Average over 5 independent seeds is reported.

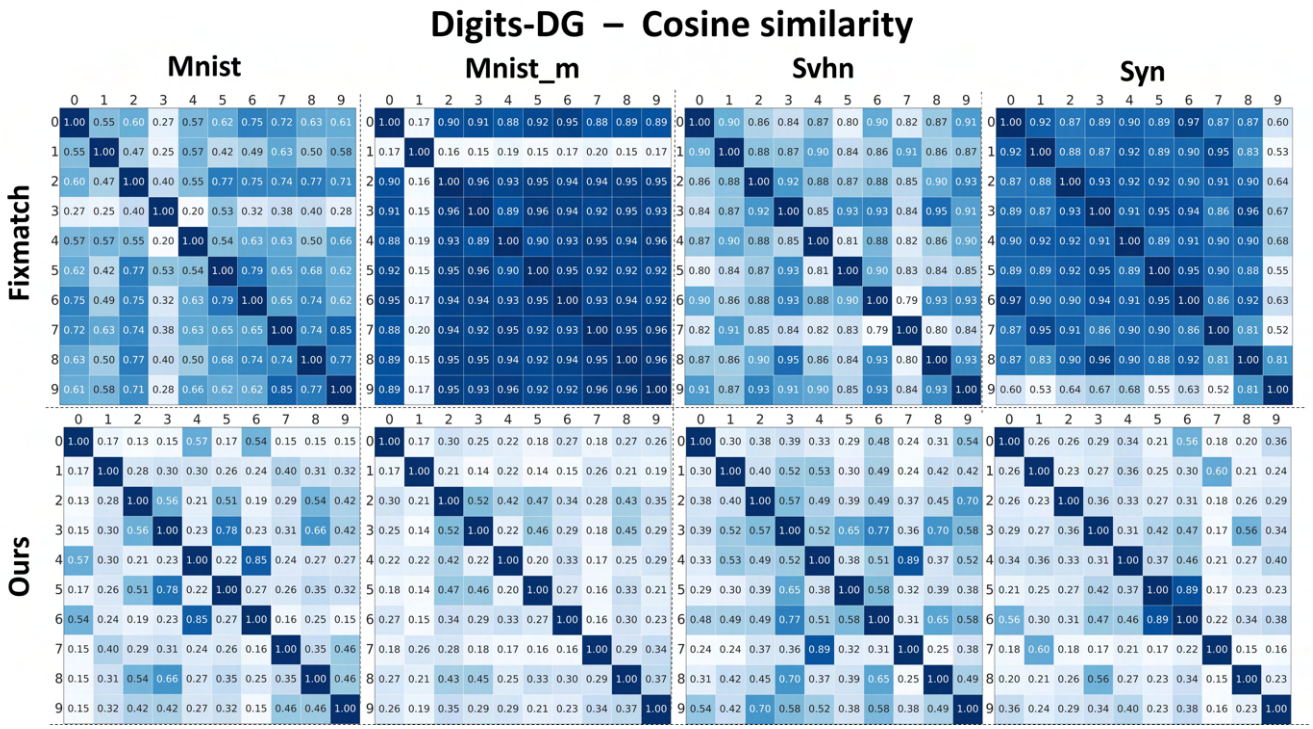


Figure 7. Comparison of cosine similarity of mean class features between Fixmatch and our method on Digits-DG.

Algorithm	**RN 50**	**RN 101**	**Vit-S/32**	**Vit-B/32**	**CLIP-B/32**
FixMatch[32]	$61.3_{\pm 0.4}$	$62.8_{\pm 0.2}$	$63.7_{\pm 0.5}$	$72.0_{\pm 0.4}$	$75.3_{\pm 0.6}$
FixM. +Ours	$\mathbf{62.1}_{\pm 0.4}$	$\mathbf{64.2}_{\pm 0.1}$	$\mathbf{64.4}_{\pm 0.3}$	$\mathbf{72.9}_{\pm 0.3}$	$\mathbf{78.9}_{\pm 0.4}$

Table 7. Results with different backbones on the Office Home dataset (10 labels per class)

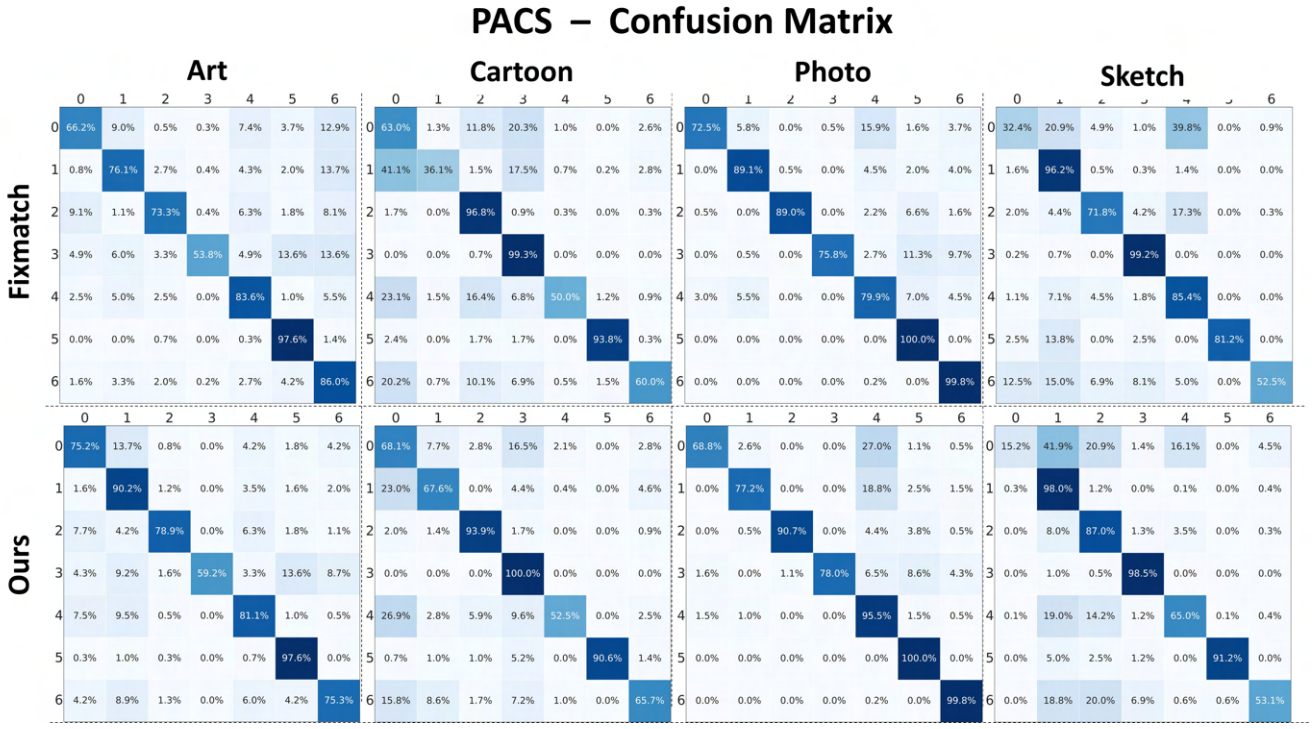


Figure 8. Confusion matrix comparison between Fixmatch and our method on PACS.

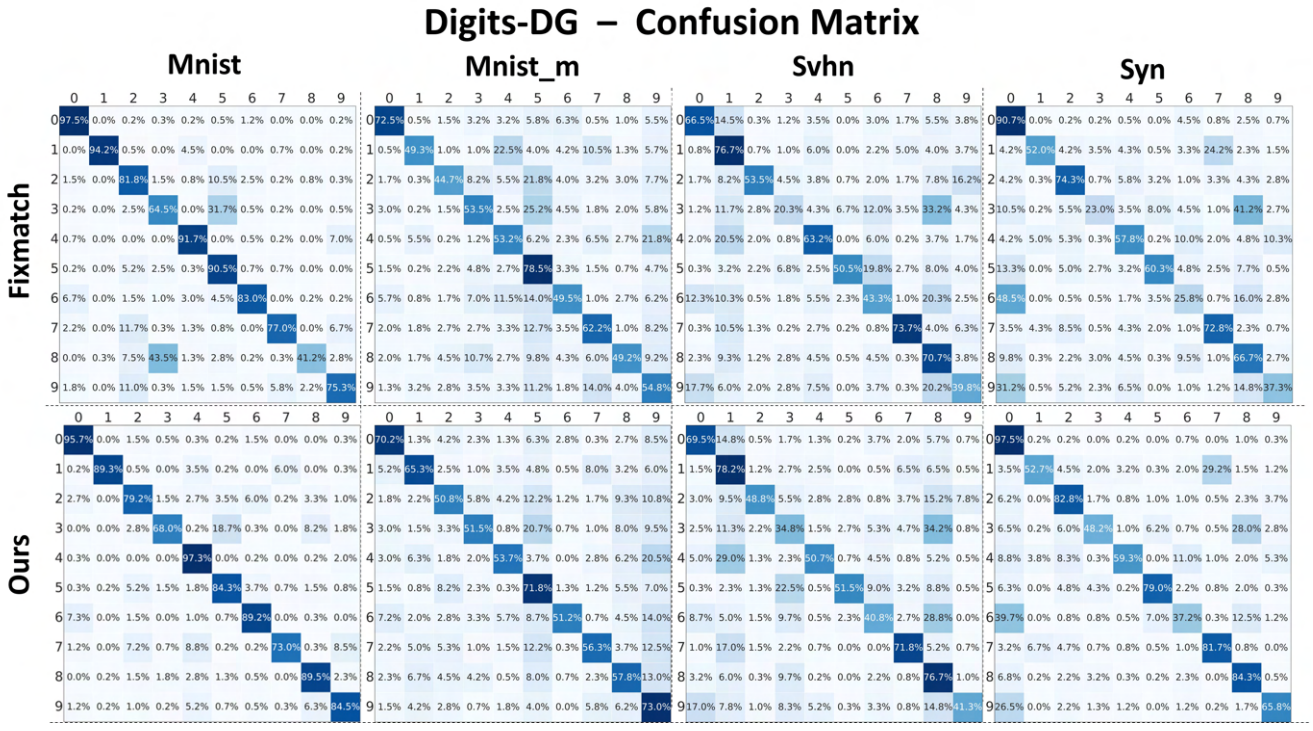


Figure 9. Confusion matrix for Fixmatch and our method on Digits-DG.

Algorithm	5	10	25	50	100
ERM[35]	$51.2_{\pm 1.0}$	$59.8_{\pm 2.5}$	$66.7_{\pm 2.2}$	$71.2_{\pm 1.9}$	$75.7_{\pm 1.6}$
FixMatch[32]	$72.8_{\pm 1.2}$	$76.6_{\pm 1.2}$	$77.6_{\pm 1.4}$	$78.7_{\pm 0.4}$	$79.4_{\pm 1.4}$
FixM.+Ours	$\mathbf{77.3}_{\pm 1.1}$	$\mathbf{78.2}_{\pm 1.2}$	$\mathbf{79.3}_{\pm 1.8}$	$\mathbf{79.6}_{\pm 1.0}$	$\mathbf{80.4}_{\pm 0.6}$

Table 8. Results with different per class labels on PACS

Algorithm	s/epoch	Overhead	GPU Mem(MB)	Overhead
FixMatch[32]	44.4	-	5595	-
StyleMatch[57]	96.38	+117.07 %	7561	+35.13 %
FixM.+Ours	52.85	+20.11 %	7647	+36.67 %

Table 9. Runtime (s/epoch) and memory (MB) overhead

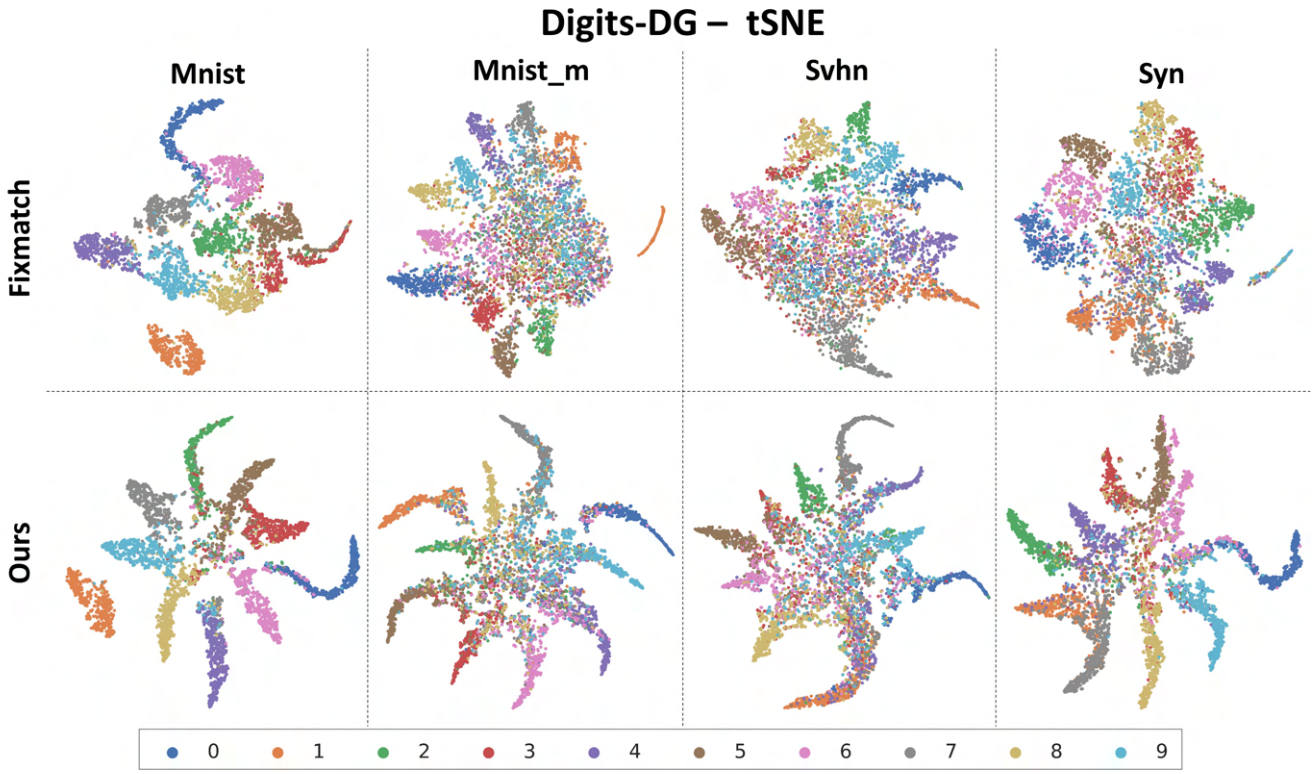


Figure 10. Feature visualization using tSNE for Fixmatch and ours on Digits-DG. Our method facilitates learning more discriminative features under various domain shifts and limited labels.