# Safe Self-Refinement for Transformer-based Domain Adaptation Supplementary Material

Tao Sun<sup>1</sup>, Cheng Lu<sup>2</sup>, Tianshuo Zhang<sup>2</sup>, Haibin Ling<sup>1</sup> <sup>1</sup>Stony Brook University, <sup>2</sup>XPeng Motors

{tao,hling}@cs.stonybrook.edu, luc@xiaopeng.com, tonyzhang2035@gmail.com

#### A. More Model and Training Details

Our implementation is based on the *timm* library<sup>1</sup>. We use ViT-B/16 [2] (*vit\_base\_patch16\_224* in *timm*) and ViT-S/16 [2] (*vit\_small\_patch16\_224* in *timm*) as the vision transformer backbones in the paper. Transformer weights are restored from the checkpoints released by official Google JAX implementation<sup>2</sup>, which are obtained by first training on ImageNet-21k [7] and then fine-tuning on Image-1k [7, 8]. The classifier head consists of a bottleneck module (Linear  $\rightarrow$  BatchNorm1d  $\rightarrow$  ReLU  $\rightarrow$  Dropout (0.5)) and a class predictor (Linear  $\rightarrow$  ReLU  $\rightarrow$  Dropout (0.5)  $\rightarrow$  Linear). The domain discriminator has the same network structure as the class predictor except having only one output.

During the training procedure, images are first resized to  $256 \times 256$  pixels, randomly flipped horizontally, and then randomly cropped and resized to  $224 \times 224$  pixels. The only exception is for VisDA-2017 [6], where center-cropping of size  $224 \times 224$  is used. During the test procedure, images are first resized to  $256 \times 256$  pixels and then center-cropped to  $224 \times 224$  pixels.

To train the model, we adopt mini-batch Stochastic Gradient Descent (SGD) with momentum of 0.9. Learning rate is scheduled as  $lr = lr_0 * (1 + 1e^{-3} \cdot i)^{-0.75}$ , where  $lr_0$  is initial learning rate, *i* is training step. The learning rate of parameters of vision transformer backbone is set to be 1/10 of *lr*. Complete hyper-parameters used for our experiments are listed in Tab. 1. Note that the same hyper-parameters are used for source-only training and baseline methods whenever applicable.

## **B.** More Analysis on Multi-layer Perturbation

Figure 1 provides additional results when adding the same amount of perturbation to each layer while not using safe training. As can be seen in the left figure, the best layer to apply perturbation varies across tasks. Besides, a layer

<sup>1</sup>https://github.com/rwightman/pytorch-image-

	Office-31	Office- Home	VisDA- 2017	DomainNet					
$ \begin{array}{c c} \alpha \\ \beta \\ \epsilon \\ T \\ L \\ batch\_size \end{array} $	64 (32 s	0. 0. 0. 10 2 ource image:	0.3 0.2 0.4 1000 4 surce images + 32 target images)						
center_crop lr <sub>0</sub> max_iters bottleneck_dim	False 0.001 10k 1024	False 0.004 20k 2048	True 0.002 20k 1024	False 0.004 40k 1024					
50 (%) 30 20 40 50 50 50 50 50 50 50 50 50 5	6 6 7 8 9 101 rmer block	100 80 560 40 40 40 40 10 80 100 80 100 80 100 80 100 10	$\begin{array}{c} Pr \rightarrow Ar \\ Pr \rightarrow Ar \\ Pr \rightarrow Cl \\ Pr \rightarrow Cl \\ 1 2 3 4 5 \\ Transform \end{array}$	6 7 8 9 1011 her block					

Table 1. Complete list of SSRT hyper-parameters used in the experiments.

Figure 1. Perturbation at different layer. <sup>†</sup>No gradient backpropagation for  $b_x^l$ .

that works for one task may fail on others. To see the importance of allowing gradient back-propagation for  $b_x^l$  (see Sec. 3.3 and Sec. 3.4 in the paper), the right figure shows that the model collapses when add perturbation to relatively deep layers while blocking the gradients of  $b_x^l$ .

Table 5 includes comparison results when adding the perturbation to raw input or a single layer ( $\{0\}$  or  $\{4\}$  or  $\{8\}$ ) in our proposed SSRT method. As can be seen, perturbing raw input performs similarly to perturbing the 0-th transformer block. Besides, perturbing any single layer degrades the performance on some adaptations tasks. In contrast, multi-layer perturbation combines their merits and obtains the best results.

models/blob/master/timm/models/vision\_transformer.py

<sup>&</sup>lt;sup>2</sup>https://github.com/google-research/vision\_transformer

## C. More Analysis on Bi-directional Self-Refinement

Table 2 provides additional results when blocking gradient back-propagation for different variables. Similar to the results listed in the paper (see Tab. 7), allowing gradient back-propagation of the teacher probabilities in KL divergence and  $b_x^l$  works better than other variants.

Table 2. Blocking gradient back-propagation for different variables. Note that  $p_x$  and  $\tilde{p}_x$  in the table only refer to the teacher probability in KL divergence. (Safe Training not applied)

	$  b_x^l$	$oldsymbol{p}_x$	$ ilde{m{p}}_x$	Cl→Ar	$Cl{\rightarrow} Pr$	$Cl{\rightarrow}Rw$
$ \begin{aligned} \omega &= 0\\ \omega &= 1\\ \omega &\sim \mathcal{B}(0.5) \end{aligned} $		× ×	×	1.61 81.17 83.68	12.71 85.00 85.69	6.08 87.28 88.04
$\begin{aligned} \omega &\sim \mathcal{B}(0.5) \\ \omega &\sim \mathcal{B}(0.5) \end{aligned}$	×			84.55 85.21	87.27 87.88	89.49 89.58

## **D.** More Analysis on Safe Training

In our method, we adopt a *Confidence Filter* to remove noisy supervisions. If it not used (*i.e.*,  $\epsilon = 0$ ), the performance may deteriorate. Table 3 shows that using Safe Training can avoid significant performance drops, making the method much safer.

Table 3. Accuracies (%) without Confidence Filter. ( $^{\dagger}$ Safe Training not applied)

	Cl→Ar	Cl→Pr	$Cl{\rightarrow}Rw$	Pr→Ar	$Pr{\rightarrow}Cl$	Pr→Rw
Baseline-B	80.06	84.12	86.67	79.52	67.03	89.44
SSRT-B <sup>†</sup>	59.33	86.98	89.74	73.92	20.30	90.59
SSRT-B	84.51	86.98	89.30	82.65	67.79	91.16

#### E. Analysis on Model's Robustness

In our proposed SSRT, we use perturbed target domain data to refine the model during the training procedure. In this section, we provide analysis on model's robustness against perturbation during the test procedure. For each testing target domain data, we follow the same way as described in the paper to add a random offset to its latent token sequence, and use the perturbed token sequence to make prediction. To analyze model's robustness against perturbation at different layers, we add perturbation to different transformer block as well as the raw input. The perturbation magnitude is controlled by a scalar  $\alpha$  as used in the paper. Figure 3 shows results (averaged over 6 random runs) on  $Pr \rightarrow Ar$  and  $clp \rightarrow pnt$ . As can be seen, our method is more robust than Baseline. Even when adding a larger amount of perturbation ( $\alpha = 0.4$ ) than seen during training, SSRT incurs less accuracy decrease.

#### F. Comparison with SSL methods

Since Unsupervised Domain Adaptation (UDA) is closely related to Semi-Supervised Learning (SSL), in this section, we compare our method with two representative techniques in SSL, *i.e.*, *Mixup* [11] and *VAT* [4].

Mixup regularizes the model to predict linearly between samples. Specifically, let  $x_1$  and  $x_2$  be two target domain data,  $p_1 = h(x_1)$  and  $p_2 = h(x_2)$  be the corresponding model predictions, Mixup first interpolates between two samples by

$$\lambda \sim Beta(\alpha_{\lambda}, \alpha_{\lambda}) \tag{1}$$

$$\boldsymbol{x}' = \lambda \boldsymbol{x}_1 + (1 - \lambda) \boldsymbol{x}_2 \tag{2}$$

$$p' = \lambda p_1 + (1 - \lambda)p_2 \tag{3}$$

Its loss function is

$$\mathcal{L}_{\text{mixup}} = \mathbb{E}_{\boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathcal{D}_t} \|h(\boldsymbol{x}') - p'\|^2$$
(4)

VAT enforces the model to predict consistently within the norm-ball neighborhood of each target data x. Its loss function is

$$\mathcal{L}_{\text{VAT}} = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{t}} \left[ \max_{\|\boldsymbol{r}\| \leq \rho} D_{KL} \left( h(\boldsymbol{x}) \| h(\boldsymbol{x} + \boldsymbol{r}) \right) \right] \quad (5)$$

We use  $\mathcal{L}_{mixup}$  and  $\mathcal{L}_{VAT}$  as the  $\mathcal{L}_{tgt}$  in our objective function. The trade-off parameter  $\beta$  is set to be 0.2 for both, same as used in our method. For Mixup,  $\alpha_{\lambda}$  is set to be 0.5. We linearly ramp up  $\beta$  to its maximum value over 1/4 of all training steps as used in [1,9]. Instead of interpolating probabilities, we interpolate unnormalized logits, as it is shown to perform slightly better. For VAT,  $\rho$  is set to be 100. Both two techniques are applied to the raw input images.

Table 4 presents results on three benchmarks using ViTbase backbone. Detailed numbers can be found in Tables 5-7. On Office-Home [10] and VisDA-2017 [6], Mixup and VAT perform better than Baseline-B, and slightly worse than ours. On DomainNet [5], VAT still works. However, for Mixup, although we tried different hyper-parameters, it is still inferior to Baseline-B. Figure 2 shows two adaptations tasks where Mixup fails.

Table 4. Comparisons with SSL methods.  $X^{\dagger}$  means averaged over all 5 tasks with X being the target domain.

	Office- Home	VisDA	Domain- Net	clp†	inf†	pnt†	qdr†	rel†	skt†
Baseline-B	81.1	85.2	38.5	50.6	25.6	44.9	11.6	57.0	41.5
Mixup-B	83.2	88.2	-	-	-	-	-	-	-
VAT-B	84.1	88.5	41.1	54.8	27.6	48.3	12.5	58.4	45.0
SSRT-B	85.4	88.8	45.2	60.0	28.2	53.3	13.7	65.3	50.4



Figure 2. Mixup with different hyperparameters. The legend for Mixup is formed as  $Mixup(\beta, \alpha_{\lambda})$ .

Figure 3. Analysis of model's robustness. The dashlines indicate true test accuracy on the target domain data. The bars show decreases of accuracies when adding perturbations to different layers during the test procedure.

Table 5. Accuracies (%) on **DomainNet**. In each sub-table, the column-wise means source domain and the row-wise means target domain. "-S/B" indicates ViT-small/base backbones, respectively.

MDD+	cln	inf	ppt	adr	ral	ekt	Δυσ	ViT-B	cln	inf	ppt	adr	ral	ekt	Δνα	Bacalina_B	cln	inf	ppt	adr	ral	ekt	Δυσ
SCDA [3]		IIII	pm	qui	ICI	SKI	Avg.	VII-D	cip	m	pm	qui	ICI	SKI	Avg.	Dasenne-D	cip	m	pin	qui	ICI	SKI	Avg.
clp	-	20.4	43.3	15.2	59.3	46.5	36.9	clp	-	27.2	53.1	13.2	71.2	53.3	43.6	clp	-	30.9	53.3	16.3	72.7	55.4	45.7
inf	32.7	-	34.5	6.3	47.6	29.2	30.1	inf	51.4	-	49.3	4.0	66.3	41.1	42.4	inf	43.0	-	40.8	7.8	56.4	35.9	36.8
pnt	46.4	19.9	-	8.1	58.8	42.9	35.2	pnt	53.1	25.6	-	4.8	70.0	41.8	39.1	pnt	55.7	28.6	-	7.4	70.5	48.3	42.1
qdr	31.1	6.6	18.0	-	28.8	22.0	21.3	qdr	30.5	4.5	16.0	-	27.0	19.3	19.5	qdr	25.5	5.2	9.7	-	15.5	17.1	14.6
rel	55.5	23.7	52.9	9.5	-	45.2	37.4	rel	58.4	29.0	60.0	6.0	-	45.8	39.9	rel	62.3	32.5	62.5	8.2	-	50.7	43.2
skt	55.8	20.1	46.5	15.0	56.7	-	38.8	skt	63.9	23.8	52.3	14.4	67.4	-	44.4	skt	66.4	30.6	58.0	18.1	70.1	-	48.6
Avg.	44.3	18.1	39.0	10.8	50.2	37.2	33.3	Avg.	51.5	22.0	46.1	8.5	60.4	40.3	38.1	Avg.	50.6	25.6	44.9	11.6	57.0	41.5	38.5
VAT-B [4]	clp	inf	pnt	qdr	rel	skt	Avg.	SSRT-B	clp	inf	pnt	qdr	rel	skt	Avg.	SSRT-B	clp	inf	pnt	qdr	rel	skt	Avg.
	1	22.1	57.1	10.5	75.0	50.0	40.0	raw input	. 1	20.7	1 .	10.0	75.0	50.0	40.2	{0}	. 1	22.2	50.7	10.6	75.0	50.7	40.2
clp	-	33.1	57.1	19.5	75.8	59.8	49.0	clp	-	32.7	60.0	19.0	75.3	59.8	49.3	clp	-	33.2	59.7	19.6	75.3	58.7	49.3
inf	48.5		45.2	9.8	55.0	57.4	39.2	inf	55.0	-	54.0	8.9	67.8	48.1	46.8	inf	54.8	-	55.5	9.3	6/./	46.1	46.3
pnt	60.0	30.9	-	7.9	/1.1	52.6	44.5	pnt	61.6	28.6	-	8.2	/1.3	35.4	45.0	pnt	61.2	29.0	-	/.1	/1.2	55.0	44.7
qdr	26.7	5.4	9.2	-	18.1	18.3	15.5	qdr	36.3	6.2	16.1	-	32.1	51.2	24.4	qdr	40.8	7.0	13.2		35.4	31.1	25.5
rel	08.7	35.3	65.0	17.6		56.8	46./	rel	69.8	35.0	60.1	12.4		59.2	48.0	rel	69.6	35.7	65./	10.7	0	58.7	48.1
SKt	10.2	33.3	49.2	17.0	12.2	-	51.7	SKT	10.5	30.5	62.3	20.0	13.2	-	51.5	SKI	69.7	32.1	62.0 50.0	19.0	12.8	-	51.1
Avg.	54.8	27.0	48.5	12.5	58.4	45.0	41.1	Avg.	38.0	20.7	51.7	13.7	63.9	50.8	44.2	Avg.	59.2	27.4	50.8	13.1	04.5	49.9	44.2
																-							
SSRT-B	clp	inf	pnt	adr	rel	skt	Avg.	SSRT-B	clp	inf	pnt	adr	rel	skt	Avg.	SSRT-B	clp	inf	pnt	adr	rel	skt	Avg.
SSRT-B {4}	clp	inf	pnt	qdr	rel	skt	Avg.	SSRT-B {8}	clp	inf	pnt	qdr	rel	skt	Avg.	SSRT-B {0,4,8}	clp	inf	pnt	qdr	rel	skt	Avg.
SSRT-B {4} clp	clp	inf 31.8	pnt 58.9	qdr 17.8	rel	skt 59.4	Avg.	SSRT-B {8} clp	clp	inf 32.4	pnt 59.0	qdr 18.6	rel	skt 59.9	Avg. 49.1	SSRT-B {0,4,8}	clp	inf 33.8	pnt 60.2	qdr 19.4	rel	skt 59.8	Avg.
SSRT-B {4} clp inf	clp - 53.5	inf 31.8	pnt 58.9 50.5	qdr 17.8 8.6	rel 75.7 67.8	skt 59.4 47.5	Avg. 48.7 45.6	SSRT-B {8} clp inf	clp 55.9	inf 32.4	pnt 59.0 54.8	qdr 18.6 7.6	rel 75.6 68.5	skt 59.9 48.2	Avg. 49.1 47.0	SSRT-B {0,4,8} clp inf	clp 55.5	inf 33.8	pnt 60.2 54.0	qdr 19.4 9.0	rel 75.8 68.2	skt 59.8 44.7	Avg. 49.8 46.3
SSRT-B {4} clp inf pnt	clp 53.5 61.3	inf 31.8 29.2	pnt 58.9 50.5	qdr 17.8 8.6 8.1	rel 75.7 67.8 71.3	skt 59.4 47.5 54.3	Avg. 48.7 45.6 44.8	SSRT-B {8} clp inf pnt	clp 55.9 61.5	inf 32.4 27.4	pnt 59.0 54.8	qdr 18.6 7.6 8.5	rel 75.6 68.5 71.4	skt 59.9 48.2 54.6	Avg. 49.1 47.0 44.7	SSRT-B {0,4,8} clp inf pnt	clp 55.5 61.7	inf 33.8 	pnt 60.2 54.0	qdr 19.4 9.0 8.4	rel 75.8 68.2 71.4	skt 59.8 44.7 55.2	Avg. 49.8 46.3 45.0
SSRT-B {4} clp inf pnt qdr	clp 53.5 61.3 42.5	inf 31.8 - 29.2 7.7	pnt 58.9 50.5 17.0	qdr 17.8 8.6 8.1	rel 75.7 67.8 71.3 23.3	skt 59.4 47.5 54.3 33.4	Avg. 48.7 45.6 44.8 24.8	SSRT-B {8} clp inf pnt qdr	clp 55.9 61.5 33.6	inf 32.4 27.4 5.7	pnt 59.0 54.8 - 11.3	qdr 18.6 7.6 8.5	rel 75.6 68.5 71.4 31.4	skt 59.9 48.2 54.6 31.8	Avg. 49.1 47.0 44.7 22.7	<b>SSRT-B</b> {0,4,8} clp inf pnt qdr	clp 55.5 61.7 42.5	inf 33.8 - 28.5 8.8	pnt 60.2 54.0 	qdr 19.4 9.0 8.4	rel 75.8 68.2 71.4 37.6	skt 59.8 44.7 55.2 33.6	Avg. 49.8 46.3 45.0 29.3
SSRT-B {4} clp inf pnt qdr rel	clp 53.5 61.3 42.5 68.7	inf 31.8 - 29.2 7.7 36.1	pnt 58.9 50.5 - 17.0 65.5	qdr 17.8 8.6 8.1 - 8.2	rel 75.7 67.8 71.3 23.3	skt 59.4 47.5 54.3 33.4 57.6	Avg. 48.7 45.6 44.8 24.8 47.2	SSRT-B {8} clp inf pnt qdr rel	clp 55.9 61.5 33.6 69.6	inf 32.4 27.4 5.7 36.2	pnt 59.0 54.8 - 11.3 65.9	qdr 18.6 7.6 8.5 - 6.9	rel 75.6 68.5 71.4 31.4	skt 59.9 48.2 54.6 31.8 58.1	Avg. 49.1 47.0 44.7 22.7 47.3	SSRT-B         {0,4,8}           clp         inf           pnt         qdr           rel         inf	clp 55.5 61.7 42.5 69.9	inf 33.8 - 28.5 8.8 37.1	pnt 60.2 54.0 - 24.2 66.0	qdr 19.4 9.0 8.4 - 10.1	rel 75.8 68.2 71.4 37.6	skt 59.8 44.7 55.2 33.6 58.9	Avg. 49.8 46.3 45.0 29.3 48.4
SSRT-B {4} clp inf pnt qdr rel skt	clp 53.5 61.3 42.5 68.7 70.1	inf 31.8 - 29.2 7.7 36.1 31.8	pnt 58.9 50.5 - 17.0 65.5 62.2	qdr 17.8 8.6 8.1 - 8.2 17.7	rel 75.7 67.8 71.3 23.3 - 73.1	skt 59.4 47.5 54.3 33.4 57.6	Avg. 48.7 45.6 44.8 24.8 47.2 51.0	SSRT-B {8} clp inf pnt qdr rel skt	clp 55.9 61.5 33.6 69.6 69.9	inf 32.4 27.4 5.7 36.2 30.9	pnt 59.0 54.8 - 11.3 65.9 62.3	qdr 18.6 7.6 8.5 - 6.9 19.8	rel 75.6 68.5 71.4 31.4 - 73.3	skt 59.9 48.2 54.6 31.8 58.1	Avg. 49.1 47.0 44.7 22.7 47.3 51.2	SSRT-B {0,4,8} clp inf pnt qdr rel skt	clp 55.5 61.7 42.5 69.9 70.6	inf 33.8 - 28.5 8.8 37.1 32.8	pnt 60.2 54.0 - 24.2 66.0 62.2	qdr 19.4 9.0 8.4 - 10.1 21.7	rel 75.8 68.2 71.4 37.6 - 73.2	skt 59.8 44.7 55.2 33.6 58.9	Avg. 49.8 46.3 45.0 29.3 48.4 52.1
SSRT-B {4} clp inf pnt qdr rel skt Avg.	clp 53.5 61.3 42.5 68.7 70.1 59.2	inf 31.8 - 29.2 7.7 36.1 31.8 27.3	pnt 58.9 50.5 - 17.0 65.5 62.2 50.8	qdr 17.8 8.6 8.1 - 8.2 17.7 12.1	rel 75.7 67.8 71.3 23.3 - 73.1 62.2	skt 59.4 47.5 54.3 33.4 57.6 - 50.4	Avg. 48.7 45.6 44.8 24.8 47.2 51.0 43.7	SSRT-B {8} clp inf pnt qdr rel skt Avg.	clp 55.9 61.5 33.6 69.6 69.9 58.1	inf 32.4 27.4 5.7 36.2 30.9 26.5	pnt 59.0 54.8 - 11.3 65.9 62.3 50.6	qdr 18.6 7.6 8.5 - 6.9 19.8 12.3	rel 75.6 68.5 71.4 31.4 - 73.3 64.0	skt 59.9 48.2 54.6 31.8 58.1 - 50.5	Avg. 49.1 47.0 44.7 22.7 47.3 51.2 43.7	SSRT-B {0,4,8} clp inf pnt qdr rel skt Avg.	clp 55.5 61.7 42.5 69.9 70.6 60.0	inf 33.8 - 28.5 8.8 37.1 32.8 28.2	pnt 60.2 54.0 24.2 66.0 62.2 53.3	qdr 19.4 9.0 8.4 - 10.1 21.7 13.7	rel 75.8 68.2 71.4 37.6 - 73.2 65.3	skt 59.8 44.7 55.2 33.6 58.9 - 50.4	Avg. 49.8 46.3 45.0 29.3 48.4 52.1 45.2
SSRT-B {4} clp inf pnt qdr rel skt Avg. ViT-S	clp 53.5 61.3 42.5 68.7 70.1 59.2 clp	inf 31.8 29.2 7.7 36.1 31.8 27.3 inf	pnt 58.9 50.5 - 17.0 65.5 62.2 50.8 pnt	qdr 17.8 8.6 8.1 - 8.2 17.7 12.1 qdr	rel 75.7 67.8 71.3 23.3 73.1 62.2 rel	skt 59.4 47.5 54.3 33.4 57.6 - 50.4 skt	Avg. 48.7 45.6 44.8 24.8 47.2 51.0 43.7 Avg.	SSRT-B {8} clp inf pnt qdr rel skt Avg. Baseline-S	clp 55.9 61.5 33.6 69.6 69.9 58.1 clp	inf 32.4 27.4 5.7 36.2 30.9 26.5 inf	pnt 59.0 54.8 - 11.3 65.9 62.3 50.6 pnt	qdr 18.6 7.6 8.5 - 6.9 19.8 12.3 qdr	rel 75.6 68.5 71.4 31.4 - 73.3 64.0 rel	skt 59.9 48.2 54.6 31.8 58.1 - 50.5 skt	Avg. 49.1 47.0 44.7 22.7 47.3 51.2 43.7 Avg.	SSRT-B {0,4,8} clp inf pnt qdr rel skt Avg. SSRT-S	clp 55.5 61.7 42.5 69.9 70.6 60.0 clp	inf 33.8 - 28.5 8.8 37.1 32.8 28.2 inf	pnt 60.2 54.0 - 24.2 66.0 62.2 53.3 pnt	qdr 19.4 9.0 8.4 - 10.1 21.7 13.7 qdr	rel 75.8 68.2 71.4 37.6 - 73.2 65.3 rel	skt 59.8 44.7 55.2 33.6 58.9 - 50.4 skt	Avg. 49.8 46.3 45.0 29.3 48.4 52.1 45.2 Avg.
SSRT-B {4} clp inf pnt qdr rel skt Avg. ViT-S clp	clp 53.5 61.3 42.5 68.7 70.1 59.2 clp	inf 31.8 - 29.2 7.7 36.1 31.8 27.3 inf 23.0	pnt 58.9 50.5 - 17.0 65.5 62.2 50.8 pnt 46.2	qdr 17.8 8.6 8.1 - 8.2 17.7 12.1 qdr 11.9	rel 75.7 67.8 71.3 23.3 - 73.1 62.2 rel 66.3	skt 59.4 47.5 54.3 33.4 57.6 - 50.4 skt 46.2	Avg. 48.7 45.6 44.8 24.8 47.2 51.0 43.7 Avg. 38.7	SSRT-B {8} clp inf pnt qdr rel skt Avg. Baseline-S clp	clp 55.9 61.5 33.6 69.6 69.9 58.1 clp	inf 32.4 5.7 36.2 30.9 26.5 inf 27.0	pnt 59.0 54.8 - 11.3 65.9 62.3 50.6 pnt 49.0	qdr 18.6 7.6 8.5 - 6.9 19.8 12.3 qdr 12.8	rel 75.6 68.5 71.4 31.4 - 73.3 64.0 rel 68.2	skt 59.9 48.2 54.6 31.8 58.1 - 50.5 skt 49.1	Avg. 49.1 47.0 44.7 22.7 47.3 51.2 43.7 Avg. 41.2	SSRT-B {0,4,8} clp inf pnt qdr rel skt Avg. SSRT-S clp	clp 55.5 61.7 42.5 69.9 70.6 60.0 clp	inf 33.8 28.5 8.8 37.1 32.8 28.2 inf 28.5	pnt 60.2 54.0 - 24.2 66.0 62.2 53.3 pnt 53.1	qdr 19.4 9.0 8.4 - 10.1 21.7 13.7 qdr 12.1	rel 75.8 68.2 71.4 37.6 - 73.2 65.3 rel 69.9	skt 59.8 44.7 55.2 33.6 58.9 - 50.4 skt 52.1	Avg. 49.8 46.3 45.0 29.3 48.4 52.1 45.2 Avg. 43.1
SSRT-B {4} clp inf pnt qdr rel skt Avg. ViT-S clp inf	clp 53.5 61.3 42.5 68.7 70.1 59.2 clp - 42.9	inf 31.8 29.2 7.7 36.1 31.8 27.3 inf 23.0	pnt 58.9 50.5 - 17.0 65.5 62.2 50.8 pnt 46.2 42.8	qdr 17.8 8.6 8.1 - 8.2 17.7 12.1 qdr 11.9 3.8	rel 75.7 67.8 71.3 23.3 - 73.1 62.2 rel 66.3 62.3	skt 59.4 47.5 54.3 33.4 57.6 - 50.4 skt 46.2 33.9	Avg.           48.7           45.6           44.8           24.8           47.2           51.0           43.7           Avg.           38.7           37.1	SSRT-B {8} clp inf pnt qdr rel skt Avg. Baseline-S clp inf	clp 55.9 61.5 33.6 69.6 69.9 58.1 clp - 41.8	inf 32.4 5.7 36.2 30.9 26.5 inf 27.0	pnt 59.0 54.8 - 11.3 65.9 62.3 50.6 pnt 49.0 43.1	qdr 18.6 7.6 8.5 - 6.9 19.8 12.3 qdr 12.8 2.7	rel 75.6 68.5 71.4 31.4 - 73.3 64.0 rel 68.2 63.0	skt 59.9 48.2 54.6 31.8 58.1 - 50.5 skt 49.1 33.0	Avg. 49.1 47.0 44.7 22.7 47.3 51.2 43.7 Avg. 41.2 36.7	SSRT-B {0,4,8} clp inf pnt qdr rel skt Avg. SSRT-S clp inf	clp 55.5 61.7 42.5 69.9 70.6 60.0 clp 47.5	inf 33.8 - 28.5 8.8 37.1 32.8 28.2 inf 28.5 -	pnt 60.2 54.0 - 24.2 66.0 62.2 53.3 pnt 53.1 49.8	qdr 19.4 9.0 8.4 - 10.1 21.7 13.7 qdr 12.1 1.5	rel 75.8 68.2 71.4 37.6 - 73.2 65.3 rel 69.9 64.9	skt 59.8 44.7 55.2 33.6 58.9 - 50.4 skt 52.1 39.7	Avg. 49.8 46.3 45.0 29.3 48.4 52.1 45.2 Avg. 43.1 40.7
SSRT-B {4} clp inf pnt qdr rel skt Avg. ViT-S clp inf pnt	clp 53.5 61.3 42.5 68.7 70.1 59.2 clp 42.9 45.2	inf 31.8 29.2 7.7 36.1 31.8 27.3 inf 23.0 - 22.2	pnt 58.9 50.5 - 17.0 65.5 62.2 50.8 pnt 46.2 42.8 -	qdr 17.8 8.6 8.1 - 8.2 17.7 12.1 qdr 11.9 3.8 3.5	rel 75.7 67.8 71.3 23.3 - 73.1 62.2 rel 66.3 62.3 66.5	skt 59.4 47.5 54.3 33.4 57.6 - 50.4 skt 46.2 33.9 35.7	Avg.           48.7           45.6           44.8           24.8           47.2           51.0           43.7           Avg.           38.7           37.1           34.6	SSRT-B {8} clp inf pnt qdr rel skt Avg. Baseline-S clp inf pnt	clp 55.9 61.5 33.6 69.6 69.9 58.1 clp - 41.8 48.8	inf 32.4 27.4 5.7 36.2 30.9 26.5 inf 27.0 - 25.7	pnt 59.0 54.8 - 11.3 65.9 62.3 50.6 pnt 49.0 43.1	qdr 18.6 7.6 8.5 - 6.9 19.8 12.3 qdr 12.8 2.7 3.1	rel 75.6 68.5 71.4 31.4 - 73.3 64.0 rel 68.2 63.0 67.0	skt 59.9 48.2 54.6 31.8 58.1 - 50.5 skt 49.1 33.0 40.8	Avg.           49.1           47.0           44.7           22.7           47.3           51.2           43.7           Avg.           41.2           36.7           37.1	SSRT-B {0,4,8} clp inf pnt qdr rel skt Avg. SSRT-S clp inf pnt	clp 55.5 61.7 42.5 69.9 70.6 60.0 clp 47.5 53.0	inf 33.8 28.5 8.8 37.1 32.8 28.2 inf 28.5 - 26.5	pnt 60.2 54.0 - 24.2 66.0 62.2 53.3 pnt 53.1 49.8 -	qdr 19.4 9.0 8.4 - 10.1 21.7 13.7 qdr 12.1 1.5 4.4	rel 75.8 68.2 71.4 37.6 - 73.2 65.3 rel 69.9 64.9 67.3	skt 59.8 44.7 55.2 33.6 58.9 - 50.4 skt 52.1 39.7 46.7	Avg. 49.8 46.3 45.0 29.3 48.4 52.1 45.2 Avg. 43.1 40.7 39.6
SSRT-B {4} clp inf pnt qdr rel skt Avg. ViT-S clp inf pnt qdr	clp 53.5 61.3 42.5 68.7 70.1 59.2 clp 42.9 45.2 19.7	inf 31.8 29.2 7.7 36.1 31.8 27.3 inf 23.0 - 22.2 3.3	pnt 58.9 50.5 - 17.0 65.5 62.2 50.8 pnt 46.2 42.8 - 7.8	qdr 17.8 8.6 8.1 - 8.2 17.7 12.1 qdr 11.9 3.8 3.5 -	rel 75.7 67.8 71.3 23.3 - 73.1 62.2 rel 66.3 62.3 66.5 14.6	skt 59.4 47.5 54.3 33.4 57.6 - 50.4 skt 46.2 33.9 35.7 12.7	Avg. 48.7 45.6 44.8 24.8 47.2 51.0 43.7 Avg. 38.7 37.1 34.6 11.6	SSRT-B {8} clp inf pnt qdr rel skt Avg. Baseline-S clp inf pnt qdr qdr	clp 55.9 61.5 33.6 69.6 69.9 58.1 clp 41.8 48.8 21.8	inf 32.4 5.7 36.2 30.9 26.5 inf 27.0 - 25.7 5.8	pnt 59.0 54.8 - 11.3 65.9 62.3 50.6 pnt 49.0 43.1 - 9.6	qdr 18.6 7.6 8.5 - 6.9 19.8 12.3 qdr 12.8 2.7 3.1 -	rel 75.6 68.5 71.4 31.4 - 73.3 64.0 rel 68.2 63.0 67.0 15.3	skt 59.9 48.2 54.6 31.8 58.1 - 50.5 skt 49.1 33.0 40.8 15.2	Avg. 49.1 47.0 44.7 22.7 47.3 51.2 43.7 Avg. 41.2 36.7 37.1 13.5	SSRT-B {0,4,8} clp inf pnt qdr rel skt Avg. SSRT-S clp inf pnt qdr	clp 55.5 61.7 42.5 69.9 70.6 60.0 clp 47.5 53.0 31.3	inf 33.8 28.5 8.8 37.1 32.8 28.2 inf 28.5 - 26.5 6.9	pnt 60.2 54.0 - 24.2 66.0 62.2 53.3 pnt 53.1 49.8 - 13.0	qdr 19.4 9.0 8.4 - 10.1 21.7 13.7 qdr 12.1 1.5 4.4 -	rel 75.8 68.2 71.4 37.6 - 73.2 65.3 rel 69.9 64.9 67.3 24.4	skt 59.8 44.7 55.2 33.6 58.9 - 50.4 skt 52.1 39.7 46.7 24.0	Avg. 49.8 46.3 45.0 29.3 48.4 52.1 45.2 Avg. 43.1 40.7 39.6 19.9
SSRT-B {4} clp inf pnt qdr rel skt Avg. ViT-S clp inf pnt qdr rel	clp 53.5 61.3 42.5 68.7 70.1 59.2 clp 42.9 45.2 19.7 50.8	inf 31.8 - 29.2 7.7 36.1 31.8 27.3 inf 23.0 - 22.2 3.3 24.2	pnt 58.9 50.5 - 17.0 65.5 62.2 50.8 pnt 46.2 42.8 - 7.8 54.2	qdr 17.8 8.6 8.1 - 8.2 17.7 12.1 qdr 11.9 3.8 3.5 - 4.6	rel 75.7 67.8 71.3 23.3 - 73.1 62.2 rel 66.3 62.3 66.5 14.6	skt 59.4 47.5 54.3 33.4 57.6 - 50.4 skt 46.2 33.9 35.7 12.7 37.3	Avg. 48.7 45.6 44.8 24.8 47.2 51.0 43.7 Avg. 38.7 37.1 34.6 11.6 34.2	SSRT-B {8} clp inf pnt qdr rel skt Avg. Baseline-S clp inf pnt qdr rel clp inf rel skt	clp 55.9 61.5 33.6 69.6 69.9 58.1 clp 41.8 48.8 21.8 54.6	inf 32.4 5.7 36.2 30.9 26.5 inf 27.0 - 25.7 5.8 28.7	pnt 59.0 54.8 - 11.3 65.9 62.3 50.6 pnt 49.0 43.1 - 9.6 57.5	qdr 18.6 7.6 8.5 - 6.9 19.8 12.3 qdr 12.8 2.7 3.1 - 3.6	rel 75.6 68.5 71.4 31.4 - 73.3 64.0 rel 68.2 63.0 67.0 15.3	skt 59.9 48.2 54.6 31.8 58.1 - 50.5 skt 49.1 33.0 40.8 15.2 41.3	Avg. 49.1 47.0 44.7 22.7 47.3 51.2 43.7 Avg. 41.2 36.7 37.1 13.5 37.1	SSRT-B {0,4,8} clp inf pnt qdr rel skt Avg. SSRT-S clp inf pnt qdr rel	clp 55.5 61.7 42.5 69.9 70.6 60.0 clp 47.5 53.0 31.3 60.0	inf 33.8 28.5 8.8 37.1 32.8 28.2 28.2 inf 28.5 - 26.5 6.9 31.2	pnt 60.2 54.0 - 24.2 66.0 62.2 53.3 pnt 53.1 49.8 - 13.0 60.5	qdr 19.4 9.0 8.4 - 10.1 21.7 13.7 qdr 12.1 1.5 4.4 - 4.6	rel 75.8 68.2 71.4 37.6 - 73.2 65.3 rel 69.9 64.9 67.3 24.4	skt 59.8 44.7 55.2 33.6 58.9 - 50.4 skt 52.1 39.7 46.7 24.0 48.5	Avg. 49.8 46.3 45.0 29.3 48.4 52.1 45.2 Avg. 43.1 40.7 39.6 19.9 41.0
SSRT-B {4} clp inf pnt qdr rel skt Avg. ViT-S clp inf pnt qdr rel skt	clp 53.5 61.3 42.5 68.7 70.1 59.2 clp 42.9 45.2 19.7 50.8 57.2	inf 31.8 - 29.2 7.7 36.1 31.8 27.3 inf 23.0 - 22.2 3.3 24.2 19.5	pnt 58.9 50.5 - 17.0 65.5 62.2 50.8 pnt 46.2 42.8 - 7.8 54.2 47.1	qdr 17.8 8.6 8.1 - 8.2 17.7 12.1 qdr 11.9 3.8 3.5 - 4.6 13.9	rel 75.7 67.8 71.3 23.3 - 73.1 62.2 rel 66.3 62.3 66.5 14.6 - 62.5	skt 59.4 47.5 54.3 33.4 57.6 - 50.4 skt 46.2 33.9 35.7 12.7 37.3 -	Avg.           48.7           45.6           44.8           24.8           47.2           51.0           43.7           Avg.           38.7           37.1           34.6           34.2           40.0	SSRT-B {8} clp inf pnt qdr rel skt Avg. Baseline-S clp inf pnt qdr rel skt	clp 55.9 61.5 33.6 69.6 69.9 58.1 clp 41.8 48.8 21.8 54.6 60.9	inf 32.4 5.7 36.2 30.9 26.5 inf 27.0 - 25.7 5.8 28.7 26.2	pnt 59.0 54.8 - 11.3 65.9 62.3 50.6 pnt 49.0 43.1 - 9.6 57.5 53.9	qdr 18.6 7.6 8.5 - 6.9 19.8 12.3 qdr 12.8 2.7 3.1 - 3.6 10.6	rel 75.6 68.5 71.4 31.4 - 73.3 64.0 rel 68.2 63.0 67.0 15.3 - 67.5	skt 59.9 48.2 54.6 31.8 58.1 - 50.5 skt 49.1 33.0 40.8 15.2 41.3	Avg. 49.1 47.0 44.7 22.7 47.3 51.2 43.7 Avg. 41.2 36.7 37.1 13.5 37.1 43.8	SSRT-B {0,4,8} clp inf pnt qdr rel skt Avg. SSRT-S clp inf pnt qdr rel skt	clp 55.5 61.7 42.5 69.9 70.6 60.0 clp 47.5 53.0 31.3 60.0 63.8	inf 33.8 - 28.5 8.8 37.1 32.8 28.2 28.2 inf 28.5 - 26.5 6.9 31.2 28.6	pnt 60.2 54.0 - 24.2 66.0 62.2 53.3 pnt 53.1 49.8 - 13.0 60.5 57.0	qdr 19.4 9.0 8.4 - 10.1 21.7 13.7 qdr 12.1 1.5 4.4 - 4.6 13.7	rel 75.8 68.2 71.4 37.6 - 73.2 65.3 rel 69.9 64.9 67.3 24.4 - 68.7	skt 59.8 44.7 55.2 33.6 58.9 - 50.4 skt 52.1 39.7 46.7 24.0 48.5	Avg. 49.8 46.3 45.0 29.3 48.4 52.1 45.2 Avg. 43.1 40.7 39.6 19.9 41.0 46.4

Table 6. Accuracies (%) on Office-Home.

Method	Ar→Cl	Ar→Pr	Ar→Rw	$Cl \rightarrow Ar$	$Cl{\rightarrow}Pr$	$Cl{\rightarrow}Rw$	Pr→Ar	$Pr{\rightarrow}Cl$	$Pr {\rightarrow} Rw$	$Rw{\rightarrow}Ar$	$Rw{\rightarrow}Cl$	$Rw{\rightarrow}Pr$	Avg.
Baseline-B Mixup-B [11]	66.96 71.32	85.74 86.66	88.07 88.82	80.06 82.45	84.12 84 79	86.67 87 58	79.52 82.90	67.03 71.68	89.44 90.77	83.64 85.46	70.15 74 36	91.17 91.37	81.05 83.18
VAT-B [4]	71.52	<b>89.39</b>	90.48	<b>86.11</b>	88.53	89.33	84.59	72.23	90.84	86.61	72.83	<b>92.48</b>	84.58
SSRT-B (ours)	75.17	88.98	91.09	85.13	88.29	89.95	85.04	74.23	91.26	85.70	78.58	91.78	85.43

Table 7. Accuracies (%) on VisDA-2017.

Method	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg.
Baseline-B	98.55	82.59	85.97	57.07	94.93	97.20	94.58	76.68	92.11	96.54	94.31	52.24	85.23
Mixup-B [11]	98.88	86.56	88.64	72.32	98.06	98.07	95.91	<b>83.00</b>	94.09	98.07	94.55	50.36	88.21
VAT-B [4]	<b>99.15</b>	<b>87.71</b>	<b>90.85</b>	67.81	<b>98.81</b>	98.17	<b>97.57</b>	76.65	92.88	<b>98.73</b>	<b>96.27</b>	<b>57.37</b> 43.13	88.50
SSRT-B (ours)	98.93	87.60	89.10	<b>84.77</b>	98.34	98.70	96.27	81.08	94.86	97.90	94.50		88.76

## G. Results with ViT-small Backbone

ViT-small is a smaller version of ViT-base by halving the number of Self-Attention Heads and token embedding dimension of ViT-base. It has fewer parameters (~22M params) than ResNet-101 (~45M params). We empirically found that it convergences much slower than ViT-base, so we double the maximum training iterations. An alternative is to pretrain on the source data first and then adapted to the target data. As can be seen from Tab. 5, our proposed SSRT-S achieves +5.1% higher accuracy than MDD+SCDA (ResNet-101 backbone) on DomainNet, despite that ViTsmall has fewer parameters than ResNet-101.

### References

- David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 2
- [2] 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. *ICLR*, 2021. 1
- [3] Shuang Li, Mixue Xie, Fangrui Lv, Chi Harold Liu, Jian Liang, Chen Qin, and Wei Li. Semantic concentration for domain adaptation. In *ICCV*, pages 9102–9111, 2021. 3
- [4] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *TPAMI*, 41(8):1979–1993, 2018. 2, 3, 4
- [5] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019. 2
- [6] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. arXiv preprint arXiv:1710.06924, 2017. 1, 2

- [7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1
- [8] Andreas Steiner, Alexander Kolesnikov, , Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.
- [9] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2
- [10] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018– 5027, 2017. 2
- [11] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: beyond empirical risk minimization. In *ICLR*, 2018. 2, 4