Supplementary material for "Cleaning Noisy Labels by Negative Ensemble Learning for Source-Free Unsupervised Domain Adaptation"

Waqar Ahmed[†‡], Pietro Morerio[†] and Vittorio Murino[†*]
[†]Pattern Analysis & Computer Vision, Istituto Italiano di Tecnologia, Genova, Italy
[‡]Dipartimento di Ingegneria Navale, Elettrica, Elettronica e delle Telecomunicazioni, University of Genova, Italy
[*] Dipartimento di Informatica, University of Verona, Italy
{waqar.ahmed, pietro.morerio, vittorio.murino}@iit.it

As mentioned in the main paper, the following supplementary material includes: (1) Performance analysis of Negative Learning against different noise distributions. (2) Noise cleaning performance of the proposed NEL method.

1. Negative Learning vs. Noise Distributions

Negative Learning (NL) is an indirect learning method in which a model is optimized to produce a lowest confidence to a randomly chosen complementary label for the given input image. Such a method produces more reliable performance in the case of a noisy label set. Nevertheless, the applicability of the existing NL method [1] is limited to a type of noise showing uniform distribution (*Symmetric-Noise*). To better highlight the limitations of existing NL method, figure 1 shows the confusion matrix for the initial noise along with the histograms of the noisy and clean sample's confidence distribution obtained after trining.

As shown in Figure 1a, when labels are initially affected by *Symmetric-Noise*, noisy samples are classified with low confidence whereas the clean samples are leaned to high confidence. Consequently, effective noise separation is achieved by NL.

For the *Asymmetric-Noise*, mimicking some of the structures of real errors [3] *i.e.,* for MNIST, mapping 2→7, 3→8, 7→1, and 5↔6, existing NL method's effectiveness degrades considerably. As can be seen in Figure 1b, noisy samples are overfitted with quite high confidence (even higher than 50%). Yet, sub-optimal separation still exists if we consider samples carrying confidence higher than 90% only(though it is not generalizable for other benchmarks where all clean samples do not carry such high confidence).

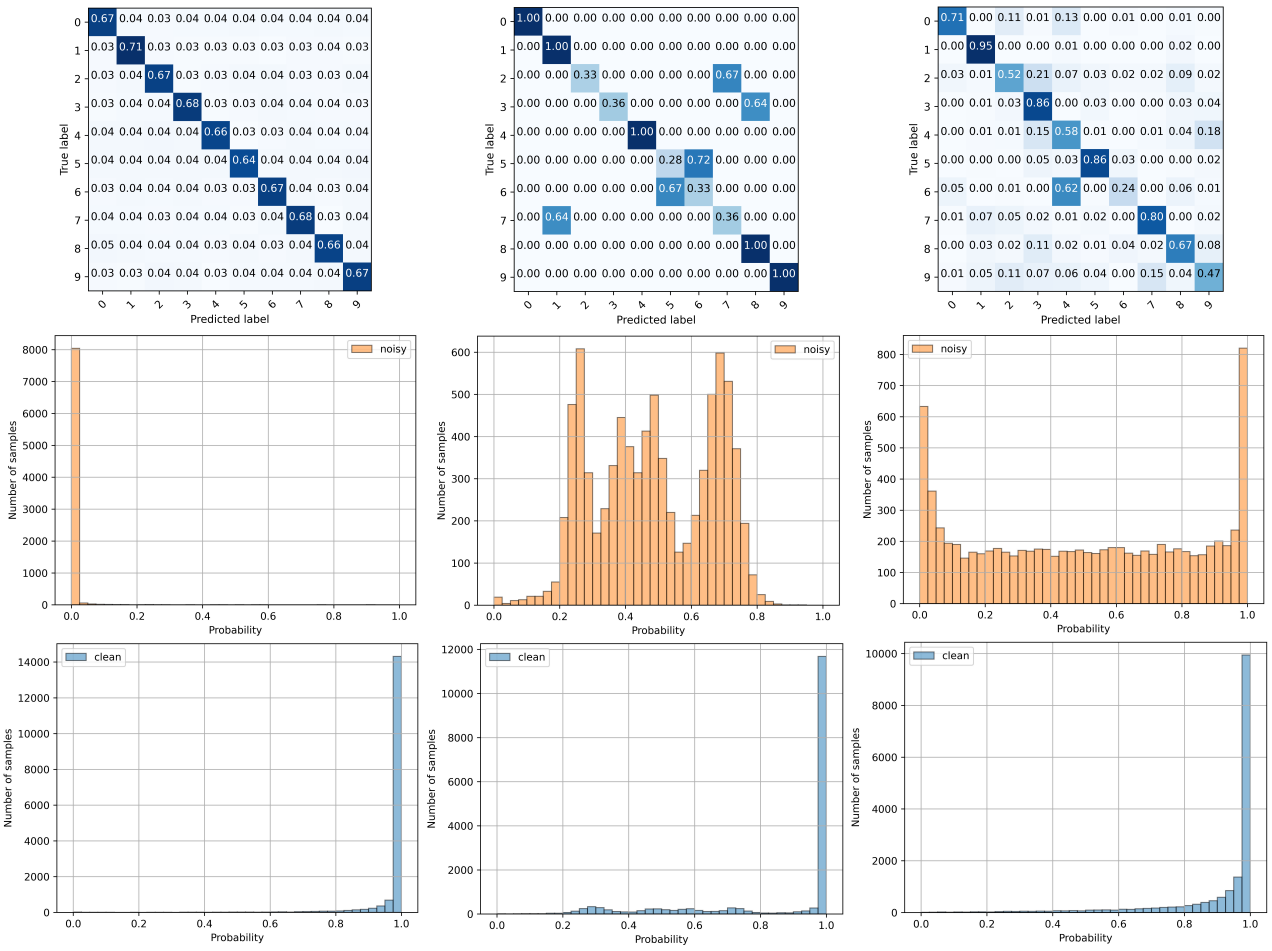
Nevertheless, noisy samples are overfitted with very high confidence when *shift-noise* associated with inferred pseudo-labels [2] is considered (Figure 1c). This happens because, in the shift-noise, inferred labels are affected by a noise skewed towards some of the classes e.g., class 3 and 4 in Figure 1c). So, instead of struggling with low confidence (the central idea of NL), noisy samples obtain higher confidence in relation to skewed (noisy) classes. Consequently, subsequent Positive Learning achieves sub-optimal performance with such noise distribution in the UDA framework.

Please note that in all cases, the *amount* of noise is same *i.e*. 32.97%, similar to the amount of shift-noise observed in the case of SVHN → MNIST UDA task in Table 1.

2. Noise Cleaning Performance of NEL

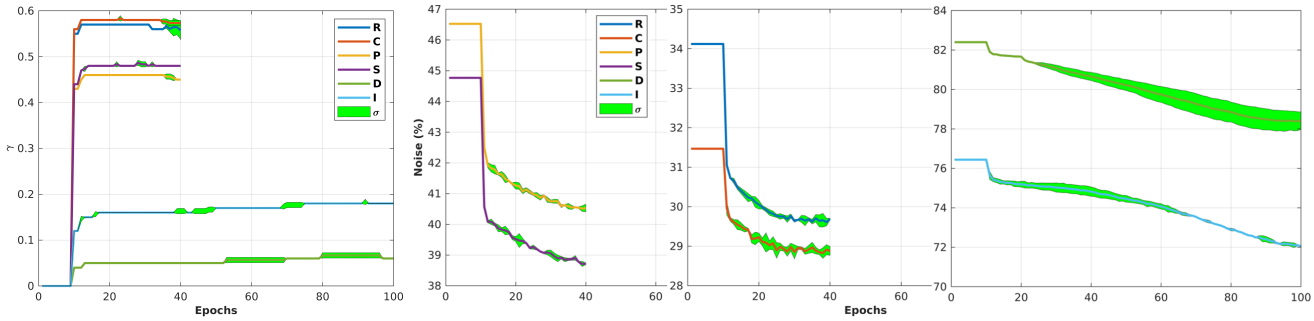
We demonstrate here the adaptive noise filtering and progressive pseudo-label refinement ability of the proposed NEL method. In Figure 2 we evaluate the robustness of our method in achieving effective pseudo-label refinement over three runs while considering the most challenging UDA task *i.e*., multi-source UDA on DomainNet. Specifically, Figure 2a shows the adaptiveness of γ threshold for different UDA tasks in which inferred pseudo-labels are affected by the various amount of noise *i.e*., starting from 31.47% up to 82.40% of shift-noise. As can be seen in Figure 2b, in each case, NEL achieves quite reasonable noise reduction throughout the training process. Such a trend is also observed in single-source UDA on *Digit5* benchmarks (see main paper). However, the only difference in the two cases is that the change in γ threshold is comparatively smoother for Digit5 which eventually takes more epochs for pseudo-label refinement (and achieves better noise reduction too).

To conclude, in Table 1-4, we summarize statistics concerning classification accuracy along with standard deviations of (i) *inferred* pseudo-labels — obtained using pre-trained source model, (ii) *refined* pseudo-labels — obtained using the proposed Negative Ensemble Learning (NEL) method, and (iii) the single-target model trained with the refined pseudo-labels — the final stage of the proposed method. In many cases, we achieve better performance with NEL only. However, for a fair comparison with existing works, the main paper compares the results achieved using a single target model only.



(a) Symmetric-Noise (b) Asymmetric-Noise (c) Shift-Noise

Figure 1: Noise filtering capability of the existing Negative Learning method [1] over various noise distributions. Column (a) Symmetric Noise, column (b) Asymmetric artificial noise [3], column (c) Shift noise [2]. First row: the confusion matrix shows how the noise is distributed in the beginning. Second row: confidence prediction for the noisy samples after training with NL. Third row: confidence prediction for the clean samples after training with NL. The amount of initial noise is same in magnitude (*i.e*. 32.97%) for all the cases.



(a) Adaptiveness of γ threshold. (b) Progressive noise reduction by pseudo-label refinement.

Figure 2: Training evolution on DomainNet. Multi-source case: for each target, the rest of the domains are considered as source. For better representation, we concatenate noise reduction trends with different scales in (b). Legend: ***C**: Clipart, **I**: Infograph, **P**: Painting, **Q**: Quickdraw, **R**: Real, **S**: Sketch, and σ: Instantaneous standard deviation of three runs.*

Source	***T***	***T***	***T***	***S***	***U***	Avg.
Target	***U***	***S***	***M***	***T***	***T***	
Inferred	86.3	34.7	63.8	67.0	70.2	64.4
Refined	99.1±0.04	62.0±0.10	97.5±0.03	99.2±0.05	99.2±0.02	91.4±0.05
NEL	97.4±0.10	61.6±0.29	95.4±0.16	99.2±0.02	99.2±0.02	90.6±0.12

(a) Single-Source UDA.

Source	***M,S,D,U***	***T,S,D,U***	***T,M,D,U***	***T,M,S,U***	***T,M,S,D***	Avg.
Target	***T***	***M***	***S***	***D***	***U***	
Inferred	98.6	69.1	52.0	40.3	88.7	69.8
Refined	98.8±0.95	94.2±0.18	84.6±0.66	87.8±0.85	98.6±0.03	92.8±0.53
NEL	99.1±0.02	95.5±0.71	89.6±0.55	90.0±0.63	97.8±0.15	94.4±0.41

(b) Multi-Source UDA.

Table 1: Results on Digit5. Legend: ***T****: MNIST,* ***S****: SVHN,* ***U****: USPS,* ***M****: MNIST-M, and* ***D****: Synthetic-Digits.*

Source	***P***			***A***			Avg.
Target (Combined)	***A,C,S***			***P,C,S***			
Inferred	37.7			57.9			47.8
Refined	57.3±1.13			73.8±0.81			65.6±0.97
Target	***A***	***C***	***S***	***P***	***C***	***S***	
NEL	80.1±0.37	76.1±1.62	25.9±0.82	96.0±0.34	82.8±1.09	49.8±0.89	68.4±0.90

(a) Multi-Target UDA. The final accuracy on each target is achieved using the same target model trained with refined pseudo-labels.

Source	***P***	***P***	***P***	***A***	***A***	***A***	Avg.
Target	***A***	***C***	***S***	***P***	***C***	***S***	
Inferred	60.9	24.8	26.5	96.0	58.1	43.9	51.7
Refined	81.1±0.44	76.9±1.36	31.7±0.98	98.2±0.19	80.9±0.77	51.2±1.04	70.0±0.80
NEL	82.6±0.83	80.5±2.66	32.3±0.68	98.4±0.03	84.3±1.50	56.1±1.27	72.4±1.16

(b) Single-Source UDA.

Source	***C,P,S***	***A,P,S***	***A,C,S***	***A,C,P***	Avg.
Target	***A***	***C***	***P***	***S***	
Inferred	78.4	77.9	95.3	64.5	79.0
Refined	89.3±0.36	87.2±0.15	98.1±0.23	83.2±0.86	89.5±0.40
NEL	90.8±0.08	89.5±0.54	98.8±0.12	85.2±0.73	91.1±0.37

(c) Multi-Source UDA.

Table 2: Results on PACS. Legend: ***A****: Art-Painting,* ***C****: Cartoon,* ***P****: Photo, and* ***S****: Sketch.*

Methods	***plane***	***bcycl***	***bus***	***car***	***horse***	***knife***	***mcycl***	***person***	***plant***	***skate***	***train***	***truck***	Avg.
Inferred	64.2	6.3	75.2	21.7	55.9	95.7	22.8	1.4	79.8	0.7	82.8	19.8	46.3
Refined	95.2	64.8	90.8	89.7	87.4	93.7	91.5	88.5	56.4	82.9	97.1	93.8	85.1
	±0.05	±0.11	±0.23	±0.27	±0.08	±0.51	±0.21	±0.66	±1.01	±0.37	±0.09	±0.15	±0.31
NEL	94.5	60.8	92.3	87.3	87.3	93.2	87.6	91.1	56.9	83.4	93.7	86.6	84.2
	±0.29	±0.31	±0.46	±0.78	±0.55	±0.02	±0.58	±0.27	±0.09	±0.44	±0.07	±0.74	±0.38

Table 3: Results on Visda-C.

Target	***C***	***I***	***P***	***Q***	***R***	***S***	Avg.
Inferred	68.5	23.6	53.5	17.6	65.9	55.2	47.4
Refined	71.1±0.11	28.0±0.05	59.5±0.12	21.6±0.44	70.4±0.04	61.3±0.03	52.0±0.13
NEL	68.3±0.15	22.1±0.17	54.7±0.15	22.8±0.45	67.3±0.92	57.1±0.27	48.7±0.35

Table 4: Multi-Source UDA results on DomainNet. Legend: ***C****: Clipart,* ***I****: Infograph,* ***P****: Painting,* ***Q****: Quickdraw,* ***R****: Real, and* ***S****: Sketch.*

References

[1] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 101–110, 2019.

[2] Pietro Morerio, Riccardo Volpi, Ruggero Ragonesi, and Vittorio Murino. Generative pseudo-label refinement for unsupervised domain adaptation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3130–3139, 2020.

[3] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.