Supplementary Material
Feature Space Perturbation: A Panacea to Enhanced Transferability Estimation

The supplementary material contains additional details and further results related to the main paper.

1. Implementation details of LDA-Based Score

This section describes the implementation process for deriving a classification score based on Linear Discriminant Analysis (LDA). LDA finds a linear combination of features that best separate the classes of data. The core optimization problem of LDA is expressed as:

$$U = \text{arg max}_U \frac{U^T \Sigma_\beta U}{U^T \Sigma_\omega U}, \tag{1}$$

where Σ_β and Σ_ω represent the between-class and within-class scatter matrices, respectively. This formula aims to project feature vectors to maximize the ratio of between-class variance to within-class variance.

To solve this optimization, we follow the methodology outlined in Ghojogh et al. (2019), which leads to the solution of the generalized eigenvalue problem:

$$U = \text{eig}((\Sigma_\omega + \epsilon I)^{-1} \Sigma_\beta), \tag{2}$$

where ϵ is a small positive scalar that ensures the non-singularity of the within-class scatter matrix Σ_ω.

In our transformed feature space, each class's features are assumed to follow a normal distribution centered at their projected class means. Bayes' theorem expresses the score function for a sample for label c as in Eq. 4. We then compute the LDA-based metric score S_{lda} using:

$$\bar{f} = U^T \mathcal{F}, \tag{3}$$

$$\delta_c = \bar{f}^T U^T \mu_c - \frac{1}{2} \mu_c^T U U^T \mu_c + \log(\frac{K_c}{K}), \tag{4}$$

$$S_{lda} = \frac{1}{K} \sum_{k=1}^{K} \frac{e^{\delta_y}}{\sum_{c=1}^{C} e^{\delta_c}}, \tag{5}$$

where y is the ground-truth class label, K is the total number of samples, and C is the number of classes. The S_{lda} represents the probability of correctly classifying a sample based on the calculated discriminative scores δ_c for each class.

2. More experimental results

This section presents experimental results that were not included in the main paper due to space limitations.

2.1. LBFT for self-supervised models

Besides vanilla fine-tuning and LFT for self-supervised models, LDA-based metrics also demonstrate improved performance in transferability estimation for LBFT. (see Table 1). The overall average values of these metrics are significantly lower compared to LFT and vanilla fine-tuning. This indicates that the previous metrics are not well-suited for LBFT fine-tuning. Observing Table 3, it's evident that the Aircraft dataset exhibits the lowest target accuracy among all datasets, suggesting that self-supervised models pre-trained with ImageNet may not be optimally suited for transfer to the Aircraft dataset. This influence is further highlighted in Table 1, where the performance of all previous metrics for the Aircraft dataset shows negative correlation scores, showing that the metric may not rank the model properly based on the feature embedding.

3. Visualization: Correlation between the metric and accuracy

Fig. 1 depicts the relationship between the ground truth target accuracy (vanilla fine-tuning) and the transferability estimation metric score across various datasets. We use the best two metrics (*i.e.*, SFDA and NCTI) to illustrate the regression plots for the original metric (depicted in pink) following the application of our feature perturbation method (depicted in blue). The shaded area indicates the 95 confidence interval. After applying our feature perturbation, the width of the shaded region in the regression plot decreases, indicating the metric score and target accuracy are more linearly correlated. Additionally, the ranking of many models shifts, bringing them closer to the shaded area, resulting in an enhancement in the weighted Kendall τ_w.

4. Downstream datasets description

We validate the effectiveness of the proposed methods on 11 standard datasets commonly adopted in transferability estimation metric evaluation. The datasets can be categorized as follows:

1. **Fine-grained classification datasets:**

   - FGVC Aircraft: This dataset contains images of various aircraft types for fine-grained classification. It consists of 100 classes with a total of 10,000 images, split into a 2:3 ratio for training and testing.
   - Stanford Cars: Comprising images of cars from different viewpoints, this dataset totals 16,185 images across various car brands and models, providing a diverse set of images for training and evaluation. The training set contains 8,144 images, while the test set contains 8,041 images.
   - Food-101: A dataset with 101,000 images categorized into 101 food classes. Each food category contains 750 training images and 250 testing images.
   - Oxford-IIIT Pets: This dataset includes 7,049 pet images belonging to 37 different pet breeds with a varying number of images per breed. The training set consists of 3,680 images, and the testing set has 3,669 images.
   - Oxford-102 Flowers: It has 102 categories with varying numbers of images per category. It comprises between 40 and 258 images per category, with 20 images sampled for training and the remaining 6,149 images for testing.

2. **Coarse-grained classification dataset:**

   - Caltech-101: A dataset with 9,146 images distributed among 101 categories. 70% of the data is sampled for the training set.
   - CIFAR-10 and CIFAR-100: These datasets contain 60,000 color images of object categories, including animals, vehicles, and everyday objects, making them suitable for general-purpose image classification. CIFAR-10 is divided into 10 distinct classes with 5,000 training images and 1,000 testing images per class. CIFAR-100 is divided into 100 distinct classes with 500 training images and 100 testing images per class.
   - VOC2007: This dataset consists of 9,963 images across 20 classes with a variety of common object classes, including people, animals, vehicles, and household items. The training set comprises 5,011 images and the remaining for testing.

3. **Scene classification dataset:**

   - SUN397: This dataset contains 397 classes, each with 1,000 scenery pictures, totaling 19,850 images. The dataset covers a wide range of scenes, including indoor/outdoor environments, and natural/urban settings.

4. **Texture classification dataset:**

   - DTD: This dataset includes 5,640 textural images categorized into 47 classes. The dataset includes high-quality images with variations in lighting, scale, and orientation, making it suitable for studying the challenges of texture recognition in real-world scenarios. Each class contains 80 training images and 40 testing images.

5. Ground Truth

The ground truth target accuracies of vanilla FT, LBFT, and LFT for supervised models are given in Table. 2, 3, 4. For self-supervised models, the ground truth target accuracies are given in Table. 5, 6, 7. For both LBFT and LFT, we follow the grid search described in [34], which selects the learning rates from $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ and weight decay parameters from $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$. Once the optimal hyper-parameters are identified, we proceed to fine-tune the pre-trained model on the designated dataset using these hyper-parameters. The resulting test accuracy serves as our benchmark. Fine-tuning is conducted on a NVIDIA A100, utilizing a batch size of 128, and all input images are resized to dimensions of 224×224. To ensure the robustness and reliability of our evaluation, we execute the code using five distinct seeds for each experiment.

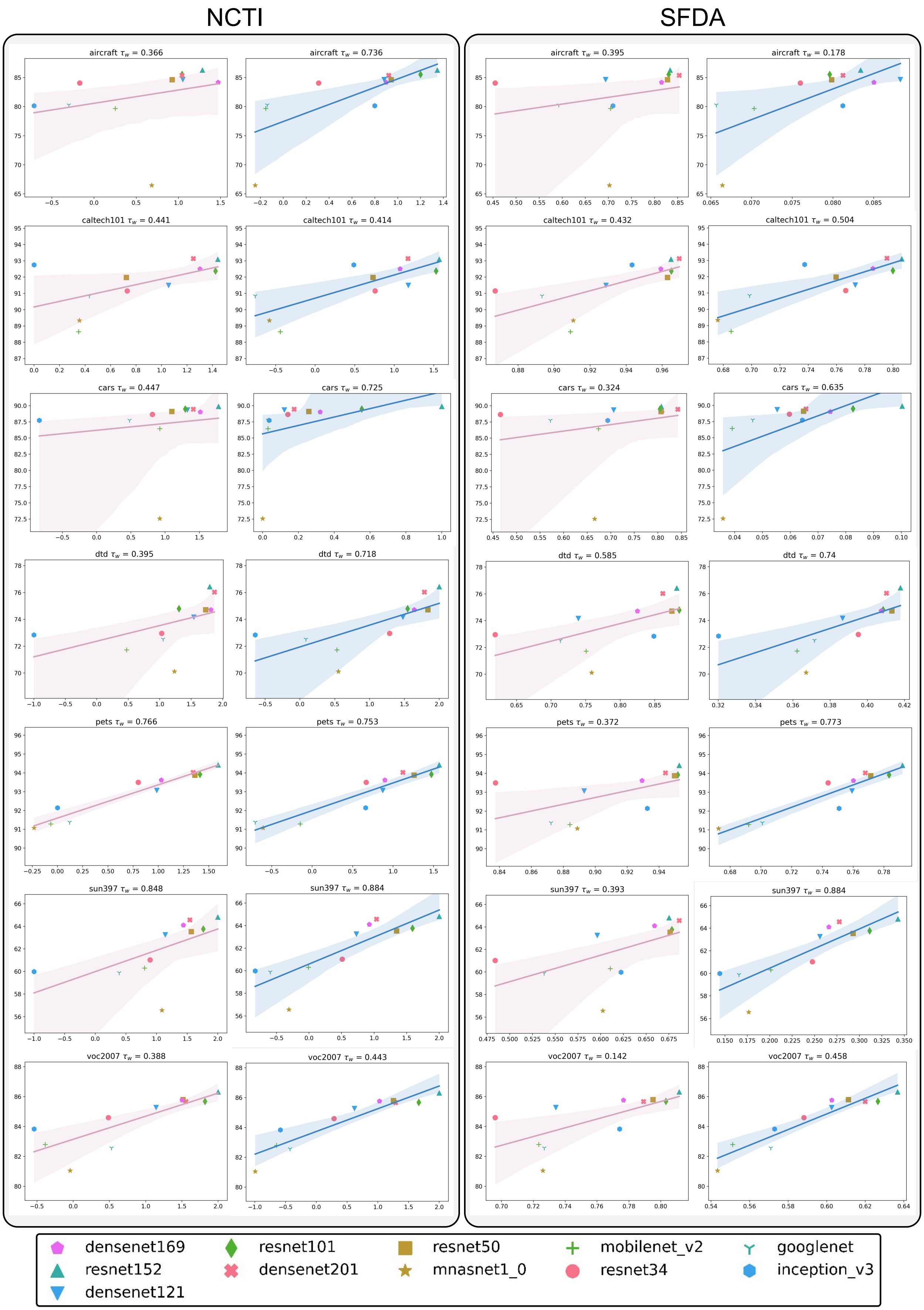

Figure 1. The figure illustrates the correlation between transferability scores and model performance (%) on the target dataset after vanilla fine-tuning (best viewed in color). Each marker denotes a distinct supervised pre-trained model. We demonstrate an enhancement for NCTI and SFDA using the weighted Kendall τ_w of our feature perturbation method (in blue) over the original method (in pink).

Table 1. Performance comparison (average weighted Kendall τ_w) for LBFT on self-supervised models. The highest performing τ_w value in each column are highlighted in bold. LDA achieves the highest overall average weighted Kendall τ_w score.

Method	Aircraft	Caltech-101	Cars	CIFAR10	CIFAR100	DTD	Flowers	Food-101	Pets	Sun397	VOC	Average
$\mathcal{N}$LEEP [24]	-0.288	0.600	0.277	**0.291**	0.326	0.849	0.283	0.603	0.683	0.099	0.184	0.355
LogME [43]	**-0.117**	0.326	0.237	-0.177	-0.183	0.836	**0.512**	**0.645**	0.708	-0.157	0.407	0.276
GBC [29]	-0.244	0.342	0.171	0.263	0.264	0.472	0.417	0.495	0.485	0.247	0.408	0.302
SFDA [33]	-0.189	0.465	0.088	-0.056	0.056	0.707	0.357	0.550	0.738	0.163	0.688	0.324
NCTI [39]	-0.194	0.610	0.046	-0.077	**0.428**	**0.895**	0.372	0.421	**0.757**	0.256	**0.738**	0.387
LDA	-0.282	**0.655**	**0.720**	0.169	0.330	0.686	0.353	0.462	0.670	**0.588**	0.681	**0.389**

Table 2. The ground truth target accuracy of vanilla fine-tuning for supervised models on 11 target datasets is sourced from [33].

	Aircraft	Caltech-101	Cars	CIFAR10	CIFAR100	DTD	Flowers	Food-101	Pets	Sun	VOC
ResNet-34	84.06	91.15	88.63	96.12	81.94	72.96	95.2	81.99	93.5	61.02	84.6
ResNet-50	84.64	91.98	89.09	96.28	82.8	74.72	96.26	84.45	93.88	63.54	85.8
ResNet-101	85.53	92.38	89.47	97.39	84.88	74.8	96.53	85.58	93.92	63.76	85.68
ResNet-152	86.29	93.1	89.88	97.53	85.66	76.44	96.86	86.28	94.42	64.82	86.32
DenseNet-121	84.66	91.5	89.34	96.45	82.75	74.18	97.02	84.99	93.07	63.26	85.28
DenseNet-169	84.19	92.51	89.02	96.77	84.26	74.72	97.32	85.84	93.62	64.1	85.77
DenseNet-201	85.38	93.14	89.44	97.02	84.88	76.04	97.1	86.71	94.03	64.57	85.67
MNet-A1	66.48	89.34	72.58	92.59	72.04	70.12	95.39	71.35	91.08	56.56	81.06
MobileNetV2	79.68	88.64	86.44	94.74	78.11	71.72	96.2	81.12	91.28	60.29	82.8
Googlenet	80.32	90.85	87.76	95.54	79.84	72.53	95.76	79.3	91.38	59.89	82.58
InceptionV3	80.15	92.75	87.74	96.18	81.49	72.85	95.73	81.76	92.14	59.98	83.84

Table 3. The ground truth target accuracy of LBFT for supervised models on 11 target datasets.

	Aircraft	Caltech-101	Cars	CIFAR10	CIFAR100	DTD	Flowers	Food-101	Pets	Sun	VOC
InceptionV3	47.98	90.25	56.6	83.76	63.46	68.99	90.76	63.77	87.78	81.6	80.88
MobileNetV2	53.33	86.78	69.83	87.06	66.59	73.19	94.95	72.24	90.41	83.26	82.03
MNet-A1	52.05	88.92	65.08	74.4	40.68	68.03	93.95	67.05	90.54	73	82.37
DenseNet-121	67.81	90.24	81.72	93.7	78.23	72.93	97.36	79.29	91.26	90.39	84.42
DenseNet-169	74.29	92.51	83.94	96.06	84.2	74.36	96.47	82.06	93.65	96.83	86.03
DenseNet-201	71.51	92.02	83.51	95.74	83.85	74.41	97.36	81.94	91.89	97.02	85.36
ResNet-34	70.94	90.42	83.07	93.94	80.74	71.54	96.42	78.04	92.84	94.84	84.39
ResNet-50	76.43	91.4	84.93	86.49	84.46	74.57	97.25	82.8	93.87	96.28	85.67
ResNet-101	75.57	91.92	85.19	96.34	85.04	74.79	96.48	83.02	93.44	97.41	85.76
ResNet-152	74.68	92.45	85.75	96.18	84.73	75.16	95.41	82.86	93.93	96.56	86.15
Googlenet	64.55	90.31	78.06	92.67	76.1	72.82	95.08	72.69	89.67	92.4	80.75

Table 4. The ground truth target accuracy of LFT for supervised models on 11 target datasets.

	Aircraft	Caltech-101	Cars	CIFAR10	CIFAR100	DTD	Flowers	Food-101	Pets	Sun	VOC
InceptionV3	28.21	88.48	27.6	69.87	46.39	61.28	83.01	46.31	85.85	63.72	77.01
MobileNetV2	42.24	87.35	49.77	76.97	57.46	67.77	92.27	62.6	89.73	73.25	80.88
MNet-A1	41.72	87.85	46.19	69.55	37.49	65.69	92.37	62.65	89.56	79.63	81.18
DenseNet-121	43.61	90.03	51.78	81.39	62.11	68.09	93.23	65.37	91.46	76.37	82.73
DenseNet-169	47.15	90.76	56.2	83.08	64.53	69.95	94.15	67.81	92.6	80.78	84.07
DenseNet-201	46.39	91.31	57.32	84.52	67.51	70.64	93.01	68.11	92.57	80.38	83.34
ResNet-34	38.19	89.8	32.04	78.61	59.43	66.7	90.71	60.56	91.27	71.96	82.46
ResNet-50	40.63	89.75	50.91	83.57	65.41	70.74	93.05	65.79	91.76	83.29	83.28
ResNet-101	41.21	89.81	50.6	85.24	67.64	69.57	92.3	66.5	92.34	75.61	83.85
ResNet-152	42.98	91.42	52.07	85.33	67.81	70.74	93.06	67.55	92.67	75.72	84.13
Googlenet	36.22	88.31	43.83	78.45	59.73	66.12	89.53	55.34	89.41	76.82	80.32

Table 5. The ground truth target accuracy of vanilla fine-tuning for self-supervised models on 11 target datasets is sourced from [33].

	Aircraft	Caltech-101	Cars	CIFAR10	CIFAR100	DTD	Flowers	Food-101	Pets	Sun	VOC
BYOL	82.10	91.90	89.83	96.98	83.86	76.37	96.80	85.44	91.48	63.69	85.13
Deepclusterv2	82.43	91.16	90.16	97.17	84.84	77.31	97.05	87.24	90.89	66.54	85.38
Infomin	83.78	80.86	86.90	96.72	70.89	73.47	95.81	78.82	90.92	57.67	81.41
InsDis	79.70	77.21	80.21	93.08	69.08	66.40	93.63	76.47	84.58	51.62	76.33
MoCov1	81.85	79.68	82.19	94.15	71.23	67.36	94.32	77.21	85.26	53.83	77.94
MoCov2	83.70	82.76	85.55	96.48	71.27	72.56	95.12	77.15	89.06	56.28	78.32
PCLv1	82.16	88.60	87.15	96.42	79.44	73.28	95.62	77.70	88.93	58.36	81.91
PCLv2	83.00	87.52	85.56	96.55	79.84	69.3	95.87	80.29	88.72	58.82	81.85
Sela-v2	85.42	90.53	89.85	96.85	84.36	76.03	96.22	86.37	89.61	65.74	85.52
SimCLRv1	80.54	90.94	89.98	97.09	84.49	73.97	95.33	82.2	88.53	63.46	83.29
SimCLRv2	81.50	88.58	88.82	96.22	78.91	74.71	95.39	82.23	89.18	60.93	83.08
SWAV	83.04	89.49	89.81	96.81	83.78	76.68	97.11	87.22	90.59	66.10	85.06

Table 6. The ground truth target accuracy of LBFT for self-supervised models on 11 target datasets.

	Aircraft	Caltech-101	Cars	CIFAR10	CIFAR100	DTD	Flowers	Food-101	Pets	Sun	VOC
BYOL	73.09	91.05	85.15	98.50	92.52	74.68	96.16	83.11	89.60	99.81	84.06
Deepclusterv2	71.16	89.62	83.26	95.95	84.62	75.21	95.87	83.29	89.82	99.79	84.59
Infomin	77.46	84.77	85.57	96.31	82.15	74.63	96.18	84.25	88.60	98.62	82.56
InsDis	70.65	76.14	79.48	94.16	77.20	70.74	91.91	79.05	80.28	97.96	76.50
MoCov1	73.16	78.40	81.40	94.35	77.97	71.49	92.24	78.70	83.05	98.02	78.10
MoCov2	75.70	85.18	84.37	96.23	82.12	73.46	95.47	82.57	87.78	97.87	81.20
PCLv1	76.60	85.63	83.82	96.94	85.98	73.03	94.78	80.82	85.83	99.13	80.87
PCLv2	76.65	85.07	84.94	97.75	87.97	72.45	95.11	82.62	87.05	99.17	81.42
Sela-v2	69.85	87.56	81.66	95.56	83.75	74.36	94.94	82.62	88.54	99.53	85.19
SimCLRv1	67.45	90.48	77.08	96.96	87.65	71.65	92.13	75.77	85.43	99.48	82.29
SimCLRv2	74.04	85.19	84.83	97.38	89.90	72.45	95.50	83.09	84.92	99.85	80.76
SWAV	71.65	88.49	82.84	95.72	83.69	75.85	95.66	83.31	87.62	99.80	84.22

Table 7. The ground truth target accuracy of LFT for self-supervised models on 11 target datasets.

	Aircraft	Caltech-101	Cars	CIFAR10	CIFAR100	DTD	Flowers	Food-101	Pets	Sun	VOC
BYOL	43.48	89.83	43.45	84.07	57.71	71.28	92.75	61.17	87.13	66.36	74.79
Deepclusterv2	47.44	89.34	56.19	79.43	55.19	72.45	93.65	68.62	87.08	81.41	80.91
Infomin	12.81	80.61	7.24	58.89	22.09	65.11	63.58	37.98	80.96	38.14	74.28
InsDis	10.93	51.26	3.82	42.81	15.65	56.33	58	27.06	50.77	31.08	52.17
MoCov1	10.88	54.23	3.32	45.01	15.68	54.41	54.56	26.89	53.03	31.01	55.92
MoCov2	11.51	78.43	5.52	54.22	24.09	64.89	59.73	34.86	73.62	34.81	70.54
PCLv1	7.46	70.13	3.90	50.70	22.68	52.23	36.81	21.12	68.08	25.84	67.99
PCLv2	13.99	82.41	8.20	69.79	32.66	65.90	69.71	36.15	75.51	39.06	72.15
Sela-v2	31.31	84.62	24.40	73.00	38.91	72.07	87.64	58.06	82.27	65.42	77.46
SimCLRv1	42.75	88.72	43.23	83.77	61.60	67.07	88.42	58.55	79.86	82.51	78.87
SimCLRv2	39.96	86.66	42.54	80.74	55.51	71.97	91.34	63.24	81.79	76.51	77.76
SWAV	43.25	87.85	45.94	75.93	47.59	74.15	92.54	66.42	85.23	77.48	79.28