Pick the Best Pre-trained Model: Towards Transferability Estimation for Medical Image Segmentation

Yuncheng Yang[1], Meng Wei[3], Junjun He[3], Jie Yang[1]✉, Jin Ye[3], and Yun Gu[1,2]✉

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China
[2] Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
[3] Shanghai AI Lab, Shanghai, China

Abstract. Transfer learning is a critical technique in training deep neural networks for the challenging medical image segmentation task that requires enormous resources. With the abundance of medical image data, many research institutions release models trained on various datasets that can form a huge pool of candidate source models to choose from. Hence, it's vital to estimate the source models' transferability (i.e., the ability to generalize across different downstream tasks) for proper and efficient model reuse. To make up for its deficiency when applying transfer learning to medical image segmentation, in this paper, we therefore propose a new **Transferability Estimation** (TE) method. We first analyze the drawbacks of using the existing TE algorithms for medical image segmentation and then design a source-free TE framework that considers both class consistency and feature variety for better estimation. Extensive experiments show that our method surpasses all current algorithms for transferability estimation in medical image segmentation. Code is available at here.

Keywords: Transferability Estimation · Model Selection · Medical Image Analysis · Deep Learning.

1 Introduction

The development of deep neural networks has greatly promoted medical imaging-based computer-aided diagnosis. Due to the large amount of learnable parameters in neural networks, sufficient annotated training samples are required for training. However, the labeling process of medical images is tedious and time-consuming. To address this problem, the common paradigm of *transfer learning*, which first pre-trains a model on upstream image datasets and then fine-tunes it on various target tasks, has been widely investigated in recent years [21,30,10]. Compared with the distributed training across multiple centers, there are no specific ethical issues or computational design of distributed/federated learning frameworks with the "pre-train-then-fine-tune" workflow.
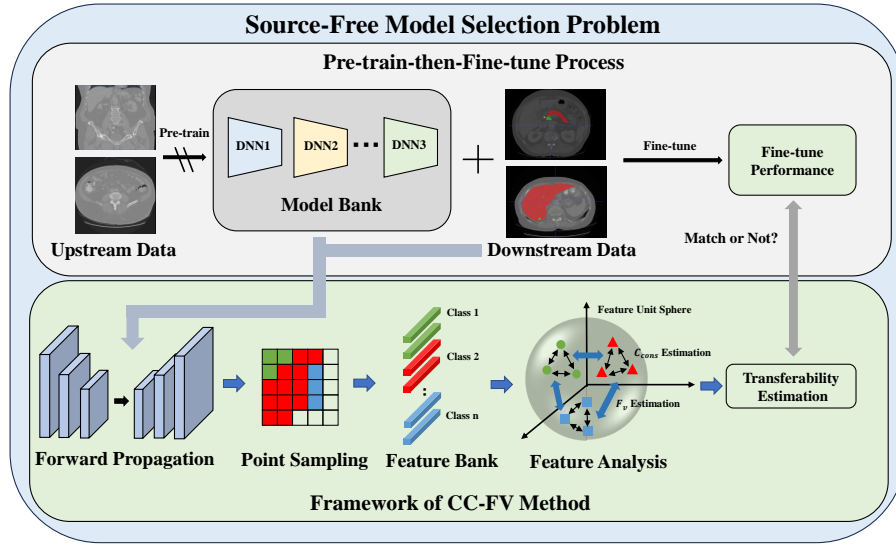
Fig. 1. Source-free model selection problem and the framework of our **C**lass **C**onsistency with **F**eature **V**ariety constraint(CC-FV) TE method. Our main goal is to predict the performance of models in the model bank after fine-tuning on downstream tasks without actually fine-tuning. Note that the upstream data are not available in our model selection process.

Previous works mainly focused on the fine-tuning strategy to effectively adapt the knowledge from the pre-trained models to target tasks [26,12,4,19]. With the increasing number of pre-trained networks provided by the community, model repositories like Hugging Face[25] and PyTorch Hub[18] enable researchers to experiment across a large number of downstream datasets and tasks. These pre-trained models require less training time and have better performance and robustness compared with the learning-from-scratch models. However, it has been observed by recent works [23] that the pre-trained models cannot always benefit the downstream tasks. When the knowledge is transferred from a less relevant source, it may not improve the performance or even negatively affect the intended outcome [24]. A brute-force method is to fine-tune a set of pre-trained models with target datasets to find the optimal one. This process is time-consuming and laborious. Existing methods also measured the task-relatedness between source and target datasets[6,7,28,22]. However, most of these works require source information available while medical images have more privacy and ethical issues and fewer datasets are publicly available than natural images.

Considering the issues mentioned above, this work focused on source-free pre-trained model selection for segmentation tasks in the medical image. As shown in Fig.1, models pre-trained by upstream data constitute the model bank. The main idea is to directly measure the transferability of the pre-trained models

without fully training based on the downstream/target dataset. Among the recent works, LEEP[15] and its variant[1,13] were developed to utilize the log-likelihood between the target labels and the predictions from the source model. LogME [27] computed evidence based on the linear parameters assumption and efficiently leverages the compatibility between features and labels. GBC [17] applied the Gaussian distribution to each class, and estimate the separability between classes as the basis for transferability estimation. TransRate[9] evaluated the transferability of models with the compactness and the completeness of embedding space. Cui et.al [5] contended that discriminability and transferability are crucial properties of representations and introduce the information bottleneck theory for transferability estimation. These methods have achieved promising performance on classification and regression tasks without fully considering the properties of medical image segmentation. First, unlike classification and regression problems that can use a single n-dimensional feature vector to represent each image, segmentation problems lack a global semantic representation, which poses difficulties for direct transferability estimation. In addition, most label-comparison-based methods[15,9,27,17] focus on the relationship between the embeddings and downstream labels without exploring the effectiveness of the features themselves. Third, medical images face severe class imbalance problems, with excessive differences between foreground and background. However, existing algorithms rarely give additional attention to the class imbalance problem. Besides, for semantic segmentation tasks, the feature pyramid is critical for the segmentation output of multi-scale objects while existing works neglect it. In our work, we propose a new method using class consistency and feature variety(CC-FV) with an efficient framework to estimate the transferability in medical image segmentation tasks. Class consistency employs the distribution of features extracted from foreground voxels of the same category in each sample to model and calculate their distance, the smaller the distance the better the result; feature diversity utilizes features sampled in the whole global feature map, and the uniformity of the feature distribution obtained by sampling is used to measure the effectiveness of the features themselves. Extensive experiments have proved the superiority of our method compared with baseline methods.

2   Methodology

2.1   Problem Formulation

In our work, a model bank $\mathbb{M}$ consisting of pre-trained models $\{M_i\}_{i=1}^{K}$ are available to be fine-tuned and evaluated with a target dataset $\mathbb{D} = \{X_j, Y_j\}_{j=1}^{N}$, where X_j is the image and Y_j is the ground truth of segmentation. After fine-tuning, the performance of M_i can be measured with the segmentation metric (e.g. Dice score), which is denoted by $\mathcal{P}_{s \to t}^{i}$ in this paper. Our work is to directly estimate the transferability score $\mathcal{T}_{s \to t}^{i}$ without fine-tuning the model on target datasets. A perfect transferability score should preserve the ordering, i.e. $\mathcal{T}_{s \to t}^{i} > \mathcal{T}_{s \to t}^{j}$ if and only if $\mathcal{P}_{s \to t}^{i} > \mathcal{P}_{s \to t}^{j}$.

2.2 Class Consistency with Feature Variety Constraint TE method

The transferability of models from a weakly related source domain to a target domain can be compromised if the domains are not sufficiently comparable [24]. This intrigues us about the question of "what kind of models are transferable". The proposed method is intuitive and straightforward: features extracted by the pre-trained model should be consistent within the class of the target dataset while representative and various globally. Therefore, **C**lass **C**onsistency and **F**eature **V**ariety are considered to estimate the transferability between models and downstream data.

Class Consistency. The pre-trained models are trained with specific pretext tasks based on the upstream dataset. Therefore, features extracted by the pre-trained models cannot perfectly distinguish the foreground and background of target data. If the features are generalizable, foreground region features will likely follow a similar distribution even without fine-tuning.

Given a pair of target data X_j and $X_{j'}$, the distribution of the features is modeled with the n-dimensional Gaussian distribution. Since the size of the foreground class varies across the cases, we therefore randomly sample the pixels/voxels of X_j and $X_{j'}$ for each class and establish the feature distribution F_j^k, $F_{j'}^k$ based on the voxels of the kth class to approximate the case-wise distribution of different classes. The class consistency between the data pair is measured by the Wasserstein distance [16] as follows:

$$\mathcal{W}_2^2(F_j^k, F_{j'}^k) = \left\| \boldsymbol{\mu}_{F_j^k} - \boldsymbol{\mu}_{F_{j'}^k} \right\|^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_{F_j^k} \right) + \mathrm{Tr}\left(\boldsymbol{\Sigma}_{F_{j'}^k} \right) - 2\,\mathrm{Tr}\left(\left(\boldsymbol{\Sigma}_{F_j^k} \boldsymbol{\Sigma}_{F_{j'}^k} \right)^{1/2} \right) \quad (1)$$

where $\mu_{F_j^k}$, $\mu_{F_{j'}^k}$ are the mean of Gaussian distribution F_j^k, $F_{j'}^k$ and $\Sigma_{F_j^k}$ and $\Sigma_{F_{j'}^k}$ are covariance matrices of F_j^k and $F_{j'}^k$. Compared to some commonly used metrics like KL-divergence or Bhattacharyya distance[17], Wasserstein distance is more stable during the computation of high-dimensional matrices because it is unnecessary to compute the determinant or inverse of a high-dimensional matrix, which can easily lead to an overflow in numerical computation. We calculate the Wasserstein distance of the distribution with voxels of the same class in a sample pair comprised of every two samples in the dataset, and obtained the following definition of class consistency C_{cons}

$$C_{cons} = \frac{1}{N(N-1)} \sum_{k=1}^{C} \sum_{i \neq j} \mathcal{W}_2(F_i^k, F_j^k) \quad (2)$$

Given that 3D medical images are computationally intensive, and prone to causing out-of-memory problems, in the sliding window inference process for each case, we do not concatenate the output of each patch into the final prediction result, but directly sample from the patched output and concatenate them into the final sampled feature matrix. In the calculation of class consistency, we only sample the foreground voxels with a pre-defined sampling number which is

proportional to the voxel number of each class in the image because of the severe class imbalance problem.

Feature Variety. Class consistency is not the only criterion for transferability estimation. As a result of learning some trivial solutions, some overfitted models have limited generalization capacity and are difficult to apply to new tasks. We believe that the essential reason for this phenomenon is that class consistency is only concerned with local homogeneity of information while neglecting the integral feature quality assessment. Hence we propose the feature variety constraint, which measures the expressiveness of the features themselves and the uniformity of their probability distribution. Highly complex features are not easily overfitted in the downstream tasks and do not collapse to cause a trivial solution.

To calculate the variety of features we need to analytically measure the properties of the feature distribution over the full feature space. Besides, to prevent overfitting and trivial features, we expect the distribution of features in the feature space to be as uniform and dispersed as possible. Therefore we employ the following hyperspherical potential energy E_s as

$$E_s\left(v\right) = \sum_{i=1}^{L} \sum_{j=1, j \neq i}^{L} e_s\left(\left\|v_i - v_j\right\|\right) = \begin{cases} \sum_{i \neq j} \left\|v_i - v_j\right\|^{-s}, & s > 0 \\ \sum_{i \neq j} \log\left(\left\|v_i - v_j\right\|^{-1}\right), & s = 0 \end{cases} \quad (3)$$

Here v is sampled feature of each image with point-wise embedding v_i and L is the length of the feature, which is also the number of sampled voxels. We randomly sample from the whole case so that the features can better express the overall representational power of the model. The feature vectors will be more widely dispersed in the unit sphere if the hyperspherical energy (HSE) is lower[3]. For the dataset with N cases, we choose $s = 1$ and the feature variety F_v is formulated as

$$F_v = \frac{1}{N} \sum_{i=1}^{N} E_s^{-1}(v) \quad (4)$$

Overall Estimation. As for semantic segmentation problems, the feature pyramid structure is critical for segmentation results[14][29]. Hence in our framework, different decoders' outputs are upsampled to the size of the output and can be used in the sliding window sampling process. Besides, we decrease the sampling ratio in the decoder layer close to the bottleneck to avoid feature redundancy. The final transferability of pre-trained model m to dataset t $\mathcal{T}_{m \to t}$ is

$$\mathcal{T}_{m \to t} = \frac{1}{D} \sum_{i=1}^{D} \log \frac{F_v^i}{C_{cons}^i} \quad (5)$$

where D is the number of decoder layers used in the estimation.

3   Experiment

3.1   Experiment on MSD Dataset

The Medical Segmentation Decathlon (MSD) [2] dataset is composed of ten different datasets with various challenging characteristics, which are widely used in the medical image analysis field. To evaluate the effectiveness of CC-FV, we conduct extensive experiments on 5 of the MSD dataset, including Task03 Liver(liver and tumor segmentation), Task06 Lung(lung nodule segmentation), Task07 Pancreas(pancreas and pancreas tumor segmentation), Task09 Spleen(spleen segmentation), and Task10 Colon(colon cancer segmentation). All of the datasets are 3D CT images. The public part of the MSD dataset is chosen for our experiments, and each dataset is divided into a training set and a test set at a scale of 80% and 20%. For each dataset, we use the other four datasets to pre-train the model and fine-tune the model on this dataset to evaluate the performance as well as the transferability using the correlation between two ranking sequences of upstream pre-trained models. We load all the pre-trained models' parameters except for the last convolutional layer and no parameters are frozen during the fine-tuning process. On top of that, we follow the nnUNet[11] with the self-configuring method to choose the pre-processing, training, and post-processing strategy. For fair comparisons, the baseline methods including TransRate [9], LogME [27], GBC [17] and LEEP [15] are also implemented. For these currently available methods, we employ the output of the layer before the final convolution as the feature map and sample it through the same sliding window as CC-FV to obtain different classes of features, which can be used for the calculation.

Fig.2 visualizes the average Dice score and the estimation value on Task 03 Liver. The TE results are obtained from the training set only. U-Net[20] and UNETR[8] are applied in the experiment and each model is pre-trained for 250k iterations and fine-tuned for 100k iterations with batch size 2 on a single NVIDIA A100 GPU. Besides, we use the model at the end of training for inference and calculate the final DSC performance on the test set. And we use weighted Kendall's τ[27] and Pearson correlation coefficient for the correlation between the TE results and fine-tuning performance. The Kendall's τ ranges from [-1, 1], and τ=1 means the rank of TE results and performance are perfectly correlated($\mathcal{T}_{s \to t}^i > \mathcal{T}_{s \to t}^j$ if and only if $\mathcal{P}_{s \to t}^i > \mathcal{P}_{s \to t}^j$). Since model selection generally picks the top models and ignores the poor performers, we assign a higher weight to the good models in the calculation, known as weighted Kendall's τ. The Pearson coefficient also ranges from [-1, 1], and measures how well the data can be described by a linear equation. The higher the Pearson coefficient, the higher the correlation between the variables. It is clear that the TE results of our method have a more positive correlation with respect to DSC performance.

Table.1 demonstrates that our method surpasses all the other methods. Most of the existing methods are inferior to ours because they are not designed for segmentation tasks with a serious class imbalance problem. Besides, these methods rely only on single-layer features and do not make good use of the hierarchical structure of the model.

Task03 Liver/Tumor Segmentation
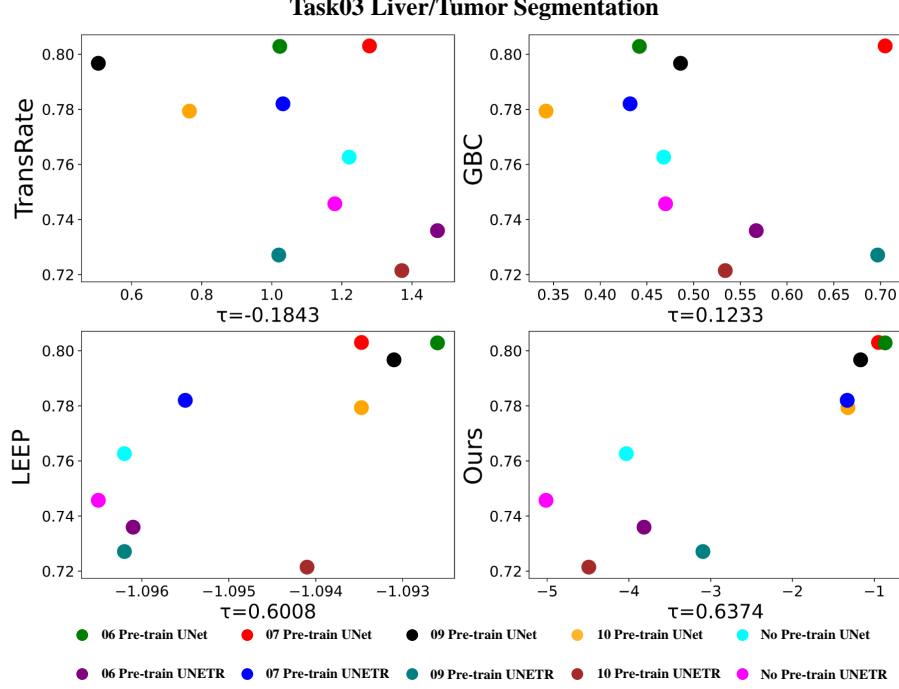
Fig. 2. Correlation between the fine-tuning performance and transferability metrics using Task03 as an example. The vertical axis represents the average Dice of the model, while the horizontal axis represents the transferability metric results. We have standardized the various metrics uniformly, aiming to observe a positive relationship between higher performance and higher transferability estimations.

Table 1. Pearson coefficient and weighted Kendall's τ for transferability estimation

Data/Method	Metrics	Task03	Task06	Task07	Task09	Task10	Avg
LogME	τ	-0.1628	-0.0988	0.3280	0.2778	-0.2348	0.0218
	pearson	0.0412	0.5713	0.3236	0.2725	-0.1674	0.2082
TransRate	τ	-0.1843	-0.1028	0.5923	0.4322	0.6069	0.2688
	pearson	-0.5178	-0.2804	0.7170	0.5573	0.7629	0.2478
LEEP	τ	0.6008	0.1658	0.2691	0.3516	0.5841	0.3943
	pearson	0.6765	-0.0073	0.7146	0.1633	0.4979	0.4090
GBC	τ	0.1233	-0.1569	0.6637	0.7611	0.6643	0.4111
	pearson	-0.2634	-0.3733	0.7948	0.7604	0.7404	0.3317
Ours CC-FV	τ	0.6374	0.0735	0.6569	0.5700	0.5550	**0.4986**
	pearson	0.8608	0.0903	0.9609	0.7491	0.8406	**0.7003**

3.2 Ablation Study

In Table.2 we analyze the different parts of our method and compare some other methods. First, we analyze the impact of class consistency C_{cons} and feature

Feature Distribution of Liver in Task03   **Feature Distribution of Tumor in Task03**
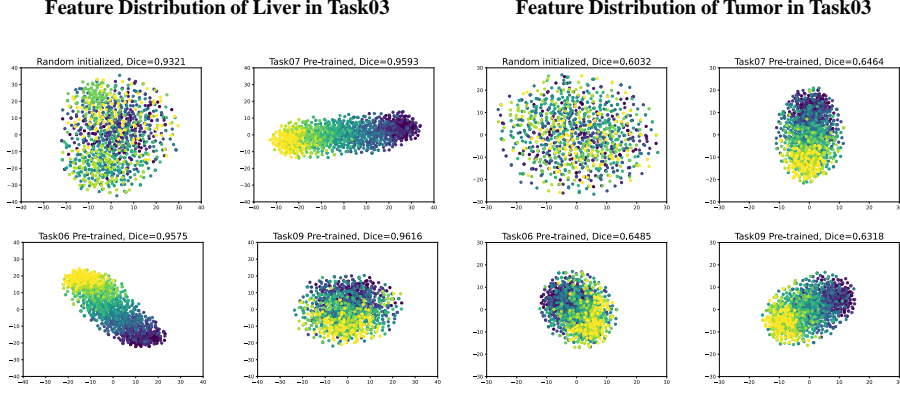
Fig. 3. Visualization of features with same labels using t-SNE. Points with different colors are from different samples. Pre-trained models tend to have a more consistent distribution within a class than the randomly initialized model and after fine-tuning they often have a better Dice performance than the randomly initialized models.

Table 2. Ablation on the effectiveness of different parts in our methods

Data/Method	Task03	Task06	Task07	Task09	Task10	Avg
Ours CC-FV	0.6374	0.0735	0.6569	0.5700	0.5550	**0.4986**
Ours w/o C_{cons}	0.1871	-0.2210	-0.2810	-0.0289	-0.2710	-0.1230
Ours w/o F_v	0.6165	0.3235	0.6054	0.2761	0.5269	0.4697
Single-scale	0.4394	0.0252	0.5336	0.5759	0.6007	0.4341
KL-divergence	-0.5658	-0.0564	0.2319	0.4628	-0.0323	0.0080
Bha-distance	0.1808	0.0723	0.2295	0.7866	0.4650	0.3468

variety F_v. Though F_v can not contribute to the final Kendall's τ directly, C_{cons} with the constraint of F_v promotes the total estimation result. Then we compare the performance of our method at single and multiple scales to prove the effectiveness of our multi-scale strategy. Finally, we change the distance metrics in class consistency estimation. KL-divergence and Bha-distance are unstable in high dimension matrics calculation and the performance is also inferior to the Wasserstein distance. Fig.3 visualize the distribution of different classes using t-SNE methods. We can easily find that with models with a pre-training process have a more compact intra-class distance and a higher fine-tuning performance.

4   Conclusion

In our work, we raise the problem of model selection for upstream and downstream transfer processes in the medical image segmentation task and analyze the practical implications of this problem. In addition, due to the ethical and privacy issues inherent in medical care and the computational load of 3D image

segmentation tasks, we design a generic framework for the task and propose a transferability estimation method based on class consistency with feature variety constraint, which outperforms existing model transferability estimation methods as demonstrated by extensive experiments.

Acknowledgement

References

1. Agostinelli, A., Uijlings, J., Mensink, T., Ferrari, V.: Transferability metrics for selecting source model ensembles. In: CVPR. pp. 7936–7946 (2022)
2. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
3. Chen, W., Yu, Z., De Mello, S., Liu, S., Alvarez, J.M., Wang, Z., Anandkumar, A.: Contrastive syn-to-real generalization. arXiv preprint arXiv:2104.02290 (2021)
4. Chen, X., Wang, S., Fu, B., Long, M., Wang, J.: Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. NIPS **32** (2019)
5. Cui, Q., Zhao, B., Chen, Z.M., Zhao, B., Song, R., Zhou, B., Liang, J., Yoshie, O.: Discriminability-transferability trade-off: an information-theoretic perspective. In: ECCV. pp. 20–37. Springer (2022)
6. Dwivedi, K., Huang, J., Cichy, R.M., Roig, G.: Duality diagram similarity: a generic framework for initialization selection in task transfer learning. In: ECCV. pp. 497–513. Springer (2020)
7. Dwivedi, K., Roig, G.: Representation similarity analysis for efficient task taxonomy & transfer learning. In: CVPR. pp. 12387–12396 (2019)
8. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: CVPR. pp. 574–584 (2022)
9. Huang, L.K., Huang, J., Rong, Y., Yang, Q., Wei, Y.: Frustratingly easy transferability estimation. In: ICML. pp. 9201–9225. PMLR (2022)
10. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI. vol. 33, pp. 590–597 (2019)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
12. Li, X., Xiong, H., Wang, H., Rao, Y., Liu, L., Chen, Z., Huan, J.: Delta: Deep learning transfer using feature map with attention for convolutional networks. arXiv preprint arXiv:1901.09229 (2019)

13. Li, Y., Jia, X., Sang, R., Zhu, Y., Green, B., Wang, L., Gong, B.: Ranking neural checkpoints. In: CVPR. pp. 2663–2673 (2021)
14. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)
15. Nguyen, C., Hassner, T., Seeger, M., Archambeau, C.: Leep: A new measure to evaluate transferability of learned representations. In: ICML. pp. 7294–7305. PMLR (2020)
16. Panaretos, V.M., Zemel, Y.: Statistical aspects of wasserstein distances. Annual review of statistics and its application **6**, 405–431 (2019)
17. Pándy, M., Agostinelli, A., Uijlings, J., Ferrari, V., Mensink, T.: Transferability estimation using bhattacharyya class separability. In: CVPR. pp. 9172–9182 (2022)
18. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
19. Reiss, T., Cohen, N., Bergman, L., Hoshen, Y.: Panda: Adapting pretrained features for anomaly detection and segmentation. In: CVPR. pp. 2806–2814 (2021)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
21. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging **35**(5), 1299–1312 (2016)
22. Tong, X., Xu, X., Huang, S.L., Zheng, L.: A mathematical framework for quantifying transferability in multi-source transfer learning. NIPS **34**, 26103–26116 (2021)
23. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: ICML. pp. 9929–9939. PMLR (2020)
24. Wang, Z., Dai, Z., Póczos, B., Carbonell, J.: Characterizing and avoiding negative transfer. In: CVPR. pp. 11293–11302 (2019)
25. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-art natural language processing. In: EMNLP. pp. 38–45 (2020)
26. Xuhong, L., Grandvalet, Y., Davoine, F.: Explicit inductive bias for transfer learning with convolutional networks. In: ICML. pp. 2825–2834. PMLR (2018)
27. You, K., Liu, Y., Wang, J., Long, M.: Logme: Practical assessment of pre-trained models for transfer learning. In: ICML. pp. 12133–12143. PMLR (2021)
28. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: CVPR. pp. 3712–3722 (2018)
29. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 2881–2890 (2017)
30. Zhou, Z., Shin, J.Y., Gurudu, S.R., Gotway, M.B., Liang, J.: Active, continual fine tuning of convolutional neural networks for reducing annotation efforts. Medical image analysis **71**, 101997 (2021)