This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

# *MetaStain*: Stain-generalizable Meta-learning for Cell Segmentation and Classification with Limited Exemplars

Aishik Konwer[1] and Prateek Prasanna[2]

[1] Department of Computer Science, Stony Brook University, NY, USA
[2] Department of Biomedical Informatics, Stony Brook University, NY, USA
akonwer@cs.stonybrook.edu, prateek.prasanna@stonybrook.edu

**Abstract.** Deep learning models excel when evaluated on test data that share similar attributes and/or distribution with the training data. However, their ability to generalize may suffer when there are discrepancies in distributions between the training and testing data i.e. domain shift. In this work, we utilize meta-learning to introduce *MetaStain*, a stain-generalizable representation learning framework that performs cell segmentation and classification in histopathology images. Owing to the designed episodical meta-learning paradigm, *MetaStain* can adapt to unseen stains and/or novel classes through finetuning even with limited annotated samples. We design a stain-aware triplet loss that clusters stain-agnostic class-specific features, as well as separates intra-stain features extracted from different classes. We also employ a consistency triplet loss to preserve the spatial correspondence between tissues under different stains. During test-time adaptation, a refined class weight generator module is optionally introduced if the unseen testing data also involves novel classes. *MetaStain* significantly outperforms state-of-the-art segmentation and classification methods on the multi-stain MIST dataset under various experimental settings.

**Keywords:** Domain Generalization · Meta Learning · Triplet loss.

## 1 Introduction

In medical imaging, the application of deep learning is often challenged by their difficulty in generalizing from one dataset to other related ones. Typically, medical images are conceptualized as points within a complex, high-dimensional, and nonlinear manifold. The inability of algorithms to consistently perform segmentation and classification tasks across different imaging techniques, patient demographics, protocols for image acquisition, and medical institutions, can largely be attributed to notable variations in the statistical properties of datasets within these manifold spaces - a phenomenon referred to as covariate shift [24]. Retraining deep learning models for each novel dataset to mitigate covariate shifts is not practically feasible in many scenarios due to the lack of rich expert-annotated training data in the novel dataset needed to re-calibrate the model parameters.

Hence, it is critical to develop techniques that exhibit robust generalization capabilities to new datasets unseen in the training phase, with minimal reliance on their annotated samples. Specifically, the problem that we intend to address here is that a model trained on multiple source domains such as Whole Slide Images (WSI) with different Immunohistochemistry (IHC) stains (Ki67, ER, HER2) [23] will find it difficult to adapt to a new target domain, for e.g. PR stained slides. The strategies of domain adaptation [7] and domain generalization [22] are designed to mitigate the effects of covariate shifts by learning domain-invariant features across both source and target datasets. Domain adaptation concentrates on generating a feature representation consistent across the distributions of both source and target domains, which ensures an adequate level of discriminative capability for the given task. On the other hand, domain generalization tackles a more complex challenge by training models across a spectrum of source domains to derive a universally applicable feature representation that is expected to be effective on target domains unseen during training. Domain generalization is a variant of transfer learning that prohibits the use of any data from the target domains during model training.

Recently, the medical vision community has delved into applying the concept of domain generalization in different ways through the usage of: (1) Nine data augmentation strategies [34] on training data to approximate the test data distribution for heart and prostrate segmentation from US and MR images, (2) an episodic meta-learning approach to improve brain tissue segmentation in T1-weighted MRI images across four medical centers [4], and (3) VAEs to identify latent subspaces to achieve patient generalization for 2D cell segmentation through domain disentanglement [14]. All the above methods have limitations: [34] applies data augmentation on a large dataset, a scenario not always practical in medical imaging where data may be scarce, aiming for domain generalization. [4] relies on meta-training and meta-test sets with similar anatomical features, which is suboptimal due to limited interactions between multiple domains. [14] operates under the assumption that each patient represents a distinct domain within the same medical center, potentially leading to similar statistical distributions between training and test data.

In this work, we propose a domain-generalization-guided meta-learning framework to perform cell segmentation and classification with limited annotated exemplars for an unseen stain, given that it encounters multiple source stains in the training stage. We employ an encoder-decoder architecture, where training is carried out in an episodic meta-learning paradigm [20] by randomly assigning source domains into meta-train and meta-test sets. This enables us to develop stain-generalizable representations that can readily adapt to perform cell segmentation and classification on an unseen stain with limited annotated exemplars in the inference stage. First, shared encoders followed by masked average pooling (MAP) generate prototype representations for each available class (IHC+, IHC-, background) in the source domain training stage. A stain-class aware hard triplet [12] loss is proposed to maximize the distance between semantic classes within a particular stain and to minimize the distance between same-class feature

clusters across multiple stains. A novel consistency-preserving loss is applied to focus on the fine-grained correspondence between tissues from different stains. Finally, finetuning is carried out on limited annotated samples of unseen target stains at test-time adaptation. For one set of experiments, a weight generator involving a Graph Attention Network (GAT) [30] is trained with support samples to learn better discriminative decision boundaries, and generate new weights to accommodate novel classes for the unseen stains at inference.

To summarize, our major contributions are as follows: 1) To the best of our knowledge, we are the first to present an episodic meta-learning training strategy in histopathology, that forces representations to be generalizable across multiple source stains during training, thus aiding our model to adapt to unseen target stains with limited annotated exemplars at inference, 2) We introduce stain-class aware hard triplet losses to produce meaningful inter- and intra-class semantic clusters and, preserve spatial consistency between tissue in different stains. 3) For one set of experiments, during test-time adaptation, we utilize class incremental learning through a GAT, that adjusts class weights to integrate novel classes at inference.

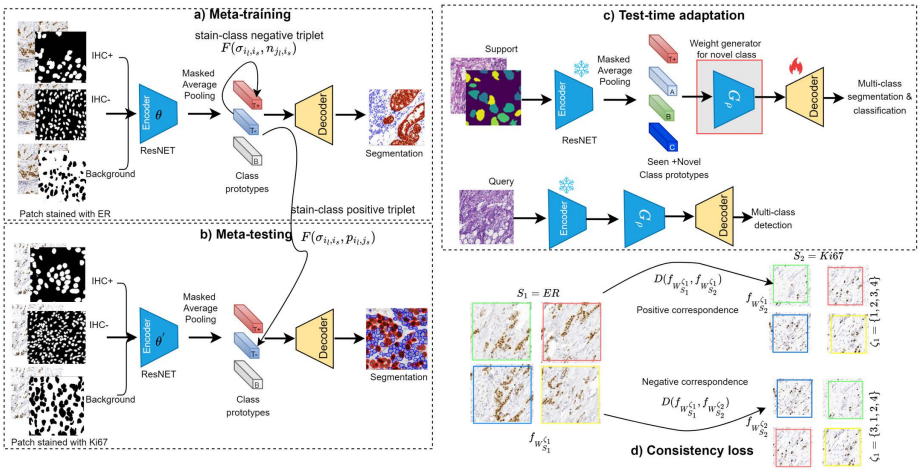


**Fig. 1. Illustration of the proposed framework.** (a-b) Meta-training and meta-testing are episodically performed on IHC-stained patch pairs. Each patch is accompanied by corresponding IHC+, IHC-, and background masks. Stain-class pair triplet loss is applied to the generated class prototypes. (c) At inference, the meta-trained model is calibrated on a small set of annotated exemplars (support) from the target stain. A weight generator is trained when the target contains novel classes. (d) Consistency triplet loss is applied for fine-grained matching between tissues across multiple stains.

## 2 Methodology

We propose a stain-generalizable meta-learning framework that can attain generalizable cell segmentation and classification performance over multiple histopathol-

**Algorithm 1** MetaStain Training

**Input: Source stains** $S = \{S_1, S_2, ..., S_n\}$, Target stain: $T$
**Initialise**: Initialise $\theta, \alpha, \beta, \omega$
**for** *iterations* $= 1, 2, \ldots$ **do**
    **Randomly Pairwise Split** 80% of source domains $\mathcal{S}$ into meta-train $S_{tr}$, and meta-test $S_{te}$. Also, construct a cohort of co-registered stains with 20% samples via image translation
    **Meta-train**: Gradients $\nabla_\theta = \mathcal{H}_\theta^{'}(S_{tr}; \theta)$
    Updated parameters: $\theta^{'} \leftarrow \theta - \alpha \nabla_\theta$
    **Meta-test**: Compute combination of Dice and weighted-CE loss with updated parameter $\theta^{'}$ as $\mathcal{G}(S_{te}; \theta^{'})$
    **Auxiliary loss**: Compute stain-class pair triplet loss $L_T$ between meta-train and meta-test and consistency preserving loss $L_{CP}$ (if co-registered cohort selected)
    **Final Model parameters**: $\theta \leftarrow \theta - \omega \frac{\partial(\mathcal{H}(S_{tr};\theta) + \beta \mathcal{G}(S_{te}; \theta - \alpha \nabla_\theta))}{\partial \theta}$
**end for**

ogy stains during the training stage, as well as quickly adapt to unseen stains and classes in the inference phase with limited annotated exemplars. Our framework consists of three sub-modules: 1) Application of meta-learning for stain generalization with multiple stains, 2) Stain-aware Triplet and Consistency preserving losses for better representation learning, and 3) Test-time adaptation with limited exemplars on unseen stains. An optional weight generator is applied in one set of experiments to accommodate, and boost the classification performance of novel classes at inference. The overall framework is illustrated in Figure 1.

### 2.1 Meta-Learning for Stain Generalization

Consider two distinct data distributions: the source, denoted as $S = \{S_1, S_2, ..., S_n\}$, and the target, represented by $T$. Both distributions are associated with the cell segmentation task and ideally should utilize identical sets of classes. However, we also explore a complex scenario of detecting novel classes at inference through class-incremental learning at test-time adaptation. Meta-Learning [6] aims to optimize a unified set of parameters, symbolized by $\theta$, through gradient descent, incorporating two key learning phases: meta-training and meta-testing. The training phase focuses solely on the source stains, $S$, while inference is carried out on the unseen target stain $T$. To facilitate this, the source $S$ is divided into two subsets: the meta-training stains (notated as $S_{tr}$) and the meta-testing domains (notated as $S_{te} = S$ - $S_{tr}$). This division is designed to replicate the scenario of shifting domains, enhancing the model's ability to adapt to the unseen target domain $T$ effectively. This learning approach is detailed in Algorithm 1 and further elaborated in Figure 1.

Consider an illustrative example involving WSI stained by three different IHC biomarkers: Ki67, ER, and PR, forming the set of source domains $S$. Additionally, there is another stain HER2, serving as the unseen test domain $T$. The proposed approach aims to train a single model parameter $\theta$ through two optimization steps. During each iteration, the images within the source stains (Ki67,

ER, PR) are randomly divided into meta-train and meta-test subsets, each consisting of images from one stain at a time. Two losses are then computed: the first loss $\mathcal{H}$ is calculated using training examples from the meta-train set, and the gradient concerning the model parameter $\theta$ is computed. The second loss $\mathcal{G}$ is computed on the meta-test set using the updated parameter $\theta^{'} = \theta - \alpha\Delta\theta$. The rationale behind introducing the second loss during the meta-test stage is that a boost in the model's performance on the meta-train set should also result in improved performance on the meta-test set. The final model parameter $\theta$ is updated by computing the gradient of the weighted combination of the two losses, $\mathcal{H}$ and $\mathcal{G}$. Specific to our scenario, both the losses are a combination of dice loss and weighted Cross-entropy [27]. This tuning ensures that the model performs well on both the meta-train and meta-test domains, preventing overfitting to any specific domain by jointly optimizing the two losses. In contrast, in a standard setup where the model is trained directly on images from all three stains (Ki67, ER, PR) without meta-learning, it may overfit to one domain while disregarding others, resulting in sub-optimal performance.

### 2.2 Stain-aware Triplet and Consistency Preserving Losses

**Stain-class pair triplet loss.** To achieve effective generalization of stain information within the embedding space, it is crucial to separate class-specific features from the same stain into distinct semantic clusters. Thus to mitigate domain discrepancies within source stains to attain desired feature clusters, we introduce a novel stain-class pair-mining technique. This domain-aware pair mining not only disperses intra-strain features according to class labels but also facilitates class-specific feature clustering independent of the stain, thereby enhancing the generalization of domain information. The two characteristics of this technique are: 1) Different classes stained with the same biomarker are pushed apart in latent space, and 2) A particular class is allowed to possess similar underlying features even though stained with different biomarkers.

Unlike conventional triplets $\{\sigma, p, n\}$ in pair-mining approaches [33], in our strategy, $p$ signifies the positive image feature with the same class but a distinct stain from $\sigma$, while $n$ represents the negative image feature with the same stain but belongs to different class from $\sigma$. The updated triplet loss can be expressed as follows: $L_T = \sum_{i \neq j} max(0, \mu_T + \mathcal{F}(\sigma_{i_l,i_s}, p_{i_l,j_s}) - \mathcal{F}(\sigma_{i_l,i_s}, n_{j_l,i_s}))$ , where $l$ and $s$ refer to the class label and stain label respectively. $\mathcal{F}$ represents the Euclidean distance between features, and $\mu_T$ is the margin of triplet loss.

**Consistency preserving loss.** The significant variability across different stains (Ki67, ER, PR, HER2) for the same tissue structure, introduces a substantial domain gap in cell segmentation and classification tasks, thereby escalating its complexity. To enhance representation learning through fine-grained matching of tissues, we employ a data augmentation strategy by shuffling patches to generate augmented triplets. We perform image-to-image translation [23] on a fraction (20%) of our dataset to generate co-registered WSIs from different stains. Next, we rearrange $n \times n$ numbered sub-patches of a random stained patch $W_{S_1}$ and

its corresponding co-registered pair ($W_{S_2}$) from a different stain, by using a permutation function $\eta(\cdot)$, resulting in shuffled versions $W^{\zeta}_{S_1} = \eta(W_{S_1}, \zeta)$ and $W^{\zeta}_{S_2} = \eta(W_{S_2}, \zeta)$, where $\zeta$ represents a random permutation of the sequence $[1, 2, ...n^2]$, dictating the rearrangement of image patches into $W^{\zeta}_{S_1}$ or $W^{\zeta}_{S_2}$

For a given co-registered patch pair from any two stains $(W_{S_1}, W_{S_2})$, the training process aims to minimize the feature distance between a patch permutation ($W^{\zeta_1}_{S_1}$) and the corresponding permutation ($\zeta_1$) of its co-registered stain patch ($W^{\zeta_1}_{S_2}$), while simultaneously maximizing the distance from a different permutation ($W^{\zeta_2}_{S_2}$). To achieve this, we formulate a triplet training objective as: $L_{CP} = max(0, \mu_c + D(f_{W^{\zeta_1}_{S_1}}, f_{W^{\zeta_1}_{S_2}}) - D(f_{W^{\zeta_1}_{S_1}}, f_{W^{\zeta_2}_{S_2}}))$, where $\mu_c$ is the margin of triplet loss.

### 2.3 Test-time adaptation with optional weight generator

We handle two scenarios at the inference stage: 1) unseen target stain with known classes, and 2) unseen target stain with novel classes. For the first scenario, we perform test-time adaptation [16] i.e. simple finetuning on some annotated support exemplars (5%-20%), and test on the rest of the target stain (query). We finetune only the decoder weights while the encoder is kept frozen.

For the second more complex scenario, we train a weight generator (involving a GAT) with some annotated support exemplars (5%-20%) and test on the rest of the target stain (query). We design a function $G_\rho$ that takes two inputs: weight vectors of known ($C_b$) and novel ($C_n$) classes, $X_{known}$ and $X_{novel}$, from the support samples passed through the frozen encoder. To facilitate information exchange among the weight vectors of $C_b + C_n$ classes, we employ GAT which is well-suited for information propagation due to its invariance to the sequence of weight vectors, allowing for the accommodation of novel classes. GAT's shared weights across nodes also enable it to adeptly handle a varying number of novel classes. The input to GAT is $X_I = \{X_{known}, X_{novel}\}$ consisting of $C_b + C_n$ weight vectors. Each weight vector corresponds to an input of a GAT node.

GAT first calculates the relationship coefficient between each pair of nodes through an inner product operation, followed by normalization with a softmax function to derive attention weights. After iterative updates of the weight vectors in the graph, the final output is the generated weight vectors for both known and novel classes as $X_{new} = G_\rho(X_I)$ with $X_I = [X_{known}; X_{novel}] \in R^{(C_b+C_n)\times d}$.

## 3 Experiment Design and Results

**Dataset.** Our experiments utilize the publicly available MIST dataset [23]. We trained our model with 5642 HER2 patches (64 WSIs), 5153 ER, 5361 Ki67, and 5139 PR patches (56 WSIs each). DeepLIIF [8] was used to generate ground truth cell masks and cell types synthetically (IHC+, IHC-). HoVer-Net [9] was used to generate ground truth for H&E images (neoplastic, non-neoplastic epithelial, inflammatory, connective, dead). In Experiment 1, we assigned each of the four

IHC stains as the unseen one during inference. In Experiment 2, during novel class segmentation, we reserve 4000 paired H&E stains, comprising 1000 samples each from HER2, ER, Ki67, and PR, for the inference stage. To implement the consistency-preserving loss, we image-translated 20% of H&E images into various stains to create co-registered ER-PR, ER-Ki67, and PR-Ki67 patches.

**Implementation Details.** For all the models, we used a ResNet50 [11] as the shared encoder in a U-Net [28] architecture. *MetaStain* is implemented in PyTorch [26] with a 48 GB Nvidia Quadro RTX 8000 GPU. All patches have dimensions of $1024 \times 1024$ and do not overlap. We trained with a batch size of one. Adam optimizer [19] was used with a linear decay scheduler and an initial learning rate of $5 \times 10^{-3}$. The hyperparameters $\alpha, \beta, \omega$ are set to 1. The loss function used is the sum of Dice [2] and weighted Cross-entropy (wCE) loss. The weights for wCE loss were 0.6 and 0.4 for IHC+ and IHC- classes. Average dice score and accuracy metrics are used for evaluation. 10 random runs have been conducted for experiments involving Ki67 and ER target stains. Average and standard deviations for Dice and Accuracy are reported in supplementary.

**Table 1.** Comparison with SOTA for IHC+ and IHC- cell segmentation and classification on Ki-67 and ER target stains. 5%-20% support samples are used for finetuning.

| Methods | Ki67 samples in inference (DSC, Acc) | | | ER samples in inference (DSC, Acc) | | |
|---|---|---|---|---|---|---|
| | 5% | 10% | 20% | 5% | 10% | 20% |
| PANet | 0.65, 0.68 | 0.68, 0.68 | 0.72, 0.71 | 0.62, 0.65 | 0.64, 0.67 | 0.69, 0.68 |
| HSNet | 0.74, 0.70 | 0.77, 0.71 | 0.81, 0.73 | 0.71, 0.68 | 0.73, 0.70 | 0.78, 0.71 |
| DCAMA | 0.77, 0.74 | 0.81, 0.76 | 0.84, 0.77 | 0.74, 0.70 | 0.76, 0.72 | 0.79, 0.74 |
| AAformer | 0.78, 0.73 | 0.81, 0.75 | 0.83, 0.76 | 0.76, 0.71 | 0.77, 0.72 | 0.80, 0.73 |
| PATNet | 0.83, 0.77 | 0.85, 0.77 | 0.87, 0.79 | 0.81, 0.74 | 0.83, 0.76 | 0.85, 0.77 |
| PMNet | 0.81, 0.76 | 0.84, 0.77 | 0.86, 0.78 | 0.79, 0.73 | 0.80, 0.74 | 0.82, 0.75 |
| MLDG-Seg | 0.80, 0.72 | 0.82, 0.75 | 0.85, 0.74 | 0.77, 0.72 | 0.79, 0.72 | 0.81, 0.76 |
| MetaMedSeg | 0.82, 0.70 | 0.80, 0.71 | 0.83, 0.70 | 0.80, 0.70 | 0.80, 0.68 | 0.79, 0.72 |
| iMAML | 0.83, 0.76 | 0.84, 0.77 | 0.86, 0.80 | 0.79, 0.77 | 0.81, 0.76 | 0.82, 0.78 |
| DCA-Net | 0.81, 0.73 | 0.83, 0.75 | 0.84, 0.77 | 0.77, 0.71 | 0.79, 0.73 | 0.80, 0.76 |
| DCAC | 0.76, 0.69 | 0.76, 0.70 | 0.77, 0.72 | 0.73, 0.66 | 0.75, 0.69 | 0.76, 0.70 |
| Ours | **0.88, 0.84** | **0.90, 0.86** | **0.91, 0.87** | **0.85, 0.84** | **0.86, 0.84** | **0.88, 0.85** |

**Experiment 1:** Our framework is evaluated on **unseen stain and known 2-class** segmentation and classification against 3 types of methods: 1) few-shot segmentation (PANet [31], DCAMA [29], and AAformer[32]) which employ various strategies including class prototype alignment, attention mechanisms, and incorporating agents in transformer aiming to improve the model's capability to generalize across diverse classes with limited examples, 2) cross-domain few-shot segmentation (HSNet [25], PATNet [21], PMNet [1]) that utilize multi-level feature correlations to perform cross-domain few-shot segmentation with minimal samples, and 3) domain generalization based segmentation methods (MLDG-Seg [18], MetaMedSeg [5], iMAML [17]) which explore variants of meta-learning for similar purpose. We also compare with DCA-Net [10] and DCA [13], which achieve domain generalization through feature augmentation and domain-invariant feature learning, rather than meta-learning. Table 1 shows consistent improvements of *MetaStain* over the second-best approach by ∼4-5% in average

Dice scores, and ∼7-9% in average accuracy for segmentation and classification of IHC+, IHC- classes in Ki67 stained samples. It should be noted that the model is finetuned with varying proportions (5%-20%) of Ki67 support exemplars at inference. Our results differ statistically significantly from SOTA, as determined by a t-test. Similar performance boosts can be observed for inference done on ER-stained patches. *MetaStain* beats the second-best approaches by 3-4% dice score in segmentation and 7-9% accuracy in cell classification when finetuned over different proportions of ER samples at inference. Results on the PR target stain are shown in the supplementary. MetaStain even surpasses [10, 13] in scenarios with 0% support samples (shown in supplementary). Qualitative results are shown in Figure 2 and supplementary.

**Experiment 2:** Our framework is evaluated on **unseen stain, novel 5-class** segmentation and classification with different proportions of annotated H&E stained support samples at inference. We explore the efficacy of each proposed component, including meta-learning (MetaL), stain-class pair loss ($L_T$), spatial consistency loss ($L_{CP}$), and weight generator($G_\rho$). We started with two types of vanilla methods: 1) Training and testing only with limited samples of target stains, 2) Training with only source stains, and finetuning on limited samples of the target domain. Both these methods performed poorly due to insufficient quantity of data or absence of parameter tuning for adaptation to a new stain. The inclusion of the triplet losses helps to form a better separation between intra-stain classes and encourages similar semantics for a class across different stains. This is reflected in the (+2-5%) boost in metrics (Table 2). Finally, $G_\rho$ helps in significantly improving average segmentation and multi-class classification accuracies of combined known and novel classes for H&E stain at inference. Ablation studies with the same variants for **unseen stain known 2-class** segmentation and classification on Ki67 samples have been reported in Table 2.

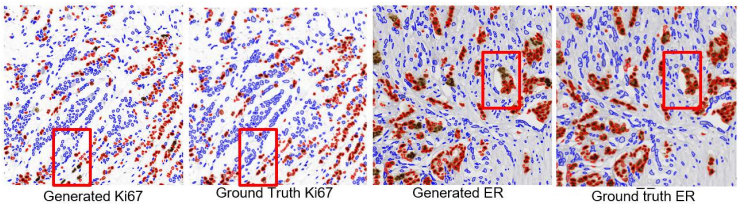


**Fig. 2.** Segmentation results of IHC+ (red) and IHC- (blue) cells on two different target stains (Ki67, ER). Regions with inconsistent segmentations are highlighted with red bounding boxes.

## 4 Conclusion

Due to poor generalizability, deep learning models often exhibit decreased performance when a domain shift exists between the training and testing data. In this work, we develop a stain-agnostic representation learning network that can

**Table 2.** Comparison with baselines for 5-class segmentation and classification in H&E target stain. Ablation results for IHC+/IHC- segmentation and classification on Ki-67.

| Methods | H&E stain in inference (avg DSC, avg Acc) | | | Ki67 stain in the inference (avg DSC, avg Acc) | | |
|---|---|---|---|---|---|---|
| | 5% | 10% | 20% | 5% | 10% | 20% |
| Vanilla target | 0.63, 0.45 | 0.64, 0.49 | 0.67,0.54 | 0.56, 0.48 | 0.59, 0.57 | 0.60, 0.61 |
| Vanilla source | 0.74, 0.59 | 0.76, 0.60 | 0.75, 0.63 | 0.62, 0.60 | 0.66, 0.63 | 0.71, 0.66 |
| MetaL | 0.79, 0.67 | 0.83, 0.69 | 0.84, 0.70 | 0.84, 0.81 | 0.87, 0.82 | 0.89, 0.83 |
| MetaL+$L_T$ | 0.83, 0.71 | 0.86, 0.73 | 0.86, 0.74 | 0.87, 0.82 | 0.89, 0.85 | 0.91, 0.84 |
| MetaL+$L_T$+$L_{CP}$ | 0.85, 0.72 | 0.87, 0.74 | 0.88, 0.75 | **0.88, 0.84** | **0.90, 0.86** | **0.91, 0.87** |
| MetaL+$L_T$+$L_{CP}$+$G_\rho$ | **0.85, 0.76** | **0.87,0.78** | **0.89, 0.81** | **-** | | |

not only generalize to multiple IHC-stain images in training but also quickly adapt to an unseen target stain with limited support samples. The meta-learned representations are further enhanced via two types of triplet losses that promote contrastive learning and spatial consistency for tissues across stains. Our method attains satisfactory multi-class segmentation and classification results when tested on unseen IHC or H&E stained WSI patches at inference. In the future, we will experiment with different proportions of co-registered samples during model training and evaluate the performance of downstream tasks involving cell morphology [3] and spatial arrangement analysis [15].

**Acknowledgments.** Research was partially supported by NIH 5R21CA258493-02.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Chen, H., Dong, Y., Lu, Z., Yu, Y., Han, J.: Pixel matching network for cross-domain few-shot segmentation. In: IEE/CVF WACV. pp. 978–987 (2024)
2. Crum, W.R., Camara, O., Hill, D.L.: Generalized overlap measures for evaluation and validation in medical image analysis. IEEE TMI **25**(11), 1451–1461 (2006)
3. Ding, R., Prasanna, P., Corredor, G., Barrera, C., Zens, P., Lu, C., Velu, P., Leo, P., Beig, N., Li, H., et al.: Image analysis reveals molecularly distinct patterns of tils in nsclc associated with treatment outcome. NPJ precision oncology **6**(1), 33 (2022)
4. Dou, Q., Coelho de Castro, D., Kamnitsas, K., Glocker, B.: Domain generalization via model-agnostic learning of semantic features. NeurIPS **32** (2019)
5. Farshad, A., Makarevich, A., Belagiannis, V., Navab, N.: Metamedseg: Volumetric meta-learning for few-shot organ segmentation. In: MICCAI Workshop on Domain Adaptation and Representation Transfer. pp. 45–55. Springer (2022)
6. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML. pp. 1126–1135. PMLR (2017)
7. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR **17**(59), 1–35 (2016)

8. Ghahremani, P., Marino, J., Dodds, R., Nadeem, S.: Deepliif: An online platform for quantification of clinical pathology slides. In: IEEE/CVF CVPR. pp. 21399–21405 (2022)
9. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.: Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. MedIA **58**, 101563 (2019)
10. Gu, R., Zhang, J., Huang, R., Lei, W., Wang, G., Zhang, S.: Domain composition and attention for unseen-domain generalizable medical image segmentation. In: MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 241–250. Springer (2021)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF CVPR. pp. 770–778 (2016)
12. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3. pp. 84–92. Springer (2015)
13. Hu, S., Liao, Z., Zhang, J., Xia, Y.: Domain and content adaptive convolution based multi-source domain generalization for medical image segmentation. IEEE TMI **42**(1), 233–244 (2022)
14. Ilse, M., Tomczak, J.M., Louizos, C., Welling, M.: Diva: Domain invariant variational autoencoders. In: MIDL. pp. 322–348. PMLR (2020)
15. Kapse, S., Pati, P., Das, S., Zhang, J., Chen, C., Vakalopoulou, M., Saltz, J., Samaras, D., Gupta, R.R., Prasanna, P.: Si-mil: Taming deep mil for self-interpretability in gigapixel histopathology. In: IEEE/CVF CVPR. pp. 11226–11237 (2024)
16. Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E.: Test-time adaptable neural networks for robust medical image segmentation. MedIA **68**, 101907 (2021)
17. Khadka, R., Jha, D., Hicks, S., Thambawita, V., Riegler, M.A., Ali, S., Halvorsen, P.: Meta-learning with implicit gradients in a few-shot setting for medical image segmentation. Computers in Biology and Medicine **143**, 105227 (2022)
18. Khandelwal, P., Yushkevich, P.: Domain generalizer: A few-shot meta learning framework for domain generalization in medical imaging. In: Domain Adaptation and Representation Transfer in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2. pp. 73–84. Springer (2020)
19. Kingma, D.P., Ba, J.L.: Adam: Amethod for stochastic optimization. In: ICLR. pp. 1–15 (2014)
20. Konwer, A., Hu, X., Bae, J., Xu, X., Chen, C., Prasanna, P.: Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation. In: IEEE/CVF ICCV. pp. 21415–21425 (2023)
21. Lei, S., Zhang, X., He, J., Chen, F., Du, B., Lu, C.T.: Cross-domain few-shot semantic segmentation. In: ECCV. pp. 73–90. Springer (2022)
22. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: IEEE ICCV. pp. 5542–5550 (2017)
23. Li, F., Hu, Z., Chen, W., Kak, A.: Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs. In: MICCAI. pp. 632–641. Springer (2023)
24. Meng, Q., Rueckert, D., Kainz, B.: Learning cross-domain generalizable features by representation disentanglement. arXiv preprint arXiv:2003.00321 (2020)
25. Min, J., Kang, D., Cho, M.: Hypercorrelation squeeze for few-shot segmentation. In: IEEE/CVF ICCV. pp. 6941–6952 (2021)
26. 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. NeurIPS **32** (2019)

27. Phan, T.H., Yamamoto, K.: Resolving class imbalance in object detection with weighted cross entropy losses. arXiv preprint arXiv:2006.01413 (2020)
28. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
29. Shi, X., Wei, D., Zhang, Y., Lu, D., Ning, M., Chen, J., Ma, K., Zheng, Y.: Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In: ECCV. pp. 151–168. Springer (2022)
30. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
31. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: ICCV (2019)
32. Wang, Y., Sun, R., Zhang, Z., Zhang, T.: Adaptive agent transformer for few-shot segmentation. In: ECCV. pp. 36–52. Springer (2022)
33. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: ECCV 2016. pp. 499–515. Springer (2016)
34. Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., Xu, D., et al.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE TMI **39**(7), 2531–2540 (2020)