Appearance Prompt Vision Transformer for Connectome Reconstruction

Rui Sun[1] , **Naisong Luo**[1] , **Yuwen Pan**[1] , **Huayu Mai**[1] ,
Tianzhu Zhang[1,2,3*] , **Zhiwei Xiong**[1,2] and **Feng Wu**[1,2]
[1]University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3]Deep Space Exploration Lab
{issunrui, lns6, panyw, mai556}@mail.ustc.edu.cn, {tzzhang, zwxiong, fengwu}@ustc.edu.cn

Abstract

Neural connectivity reconstruction aims to understand the function of biological reconstruction and promote basic scientific research. The intricate morphology and densely intertwined branches makes it an extremely challenging task. Most previous best-performing methods adopt affinity learning or metric learning. Nevertheless, they either neglect to model explicit voxel semantics caused by implicit optimization or are hysteresis to spatial information. Furthermore, the inherent locality of 3D CNNs limits modeling long-range dependencies, leading to sub-optimal results. In this work, we propose a coherent and unified Appearance Prompt Vision Transformer (APViT) to integrate affinity and metric learning to exploit the complementarity by learning long-range spatial dependencies. The proposed APViT enjoys several merits. First, the extension continuity-aware attention module aims at constructing hierarchical attention customized for neuron extensibility and slice continuity to learn instance voxel semantic context from a global perspective and utilize continuity priors to enhance voxel spatial awareness. Second, the appearance prompt modulator is responsible for leveraging voxel-adaptive appearance knowledge conditioned on affinity rich in spatial information to instruct instance voxel semantics, exploiting the potential of affinity learning to complement metric learning. Extensive experimental results on multiple challenging benchmarks demonstrate that our APViT achieves consistent improvements with huge flexibility under the same post-processing strategy.

1 Introduction

Neural connectivity reconstruction is a fundamental task to understand the function of biological reconstruction, which can widely promote basic scientific research including electrophysiology [Ascoli, 2002], cellular physiology [Donohue and Ascoli, 2011], genetics [Livet and Weissman, 2007], etc. 3D electron microscopy (EM) is the only available imaging instrument with the sufficient resolution to visualize and reconstruct dense neural morphology without ambiguity. However, at this resolution, even moderately small neural circuits yield numerous neuron numbers (e.g., typically hundreds of neuron instances in a single megapixel image [Meirovitch *et al.*, 2019a]) that are prohibitively laborious for human manual annotation (e.g., normally the human labor required to reconstruct a 100^3 μm^3 volume is at more than 100,000 hours [Berning *et al.*, 2015]). Recently, considerable works [Funke *et al.*, 2018; Januszewski *et al.*, 2018; Meirovitch *et al.*, 2019b] have turned their attention to deep neural networks in the pursuit of automatic neural connectivity reconstruction. Since all neuron instances are of the same type (i.e., biological cells), with intricate morphology and densely intertwined branches, how to fully probe discriminative information to perform accurate neuron reconstruction is thus extremely challenging.

To tackle the neural connectivity reconstruction problem, existing methods can be roughly categorized as object tracking based and boundary detection based paradigms. In the object tracking based paradigm [Januszewski *et al.*, 2018; Meirovitch *et al.*, 2016], 3D recurrent convolutional neural

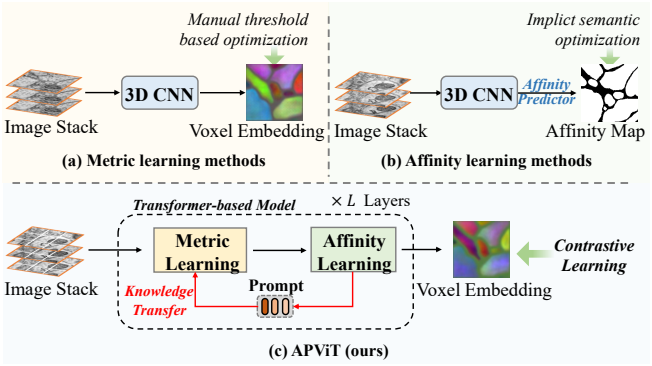

Figure 1: Different learning formulation for neural connectivity reconstruction. (a) Metric learning methods with manual threshold based optimization. (b) Affinity learning methods with an intuitive yet implicit optimization. (c) Our proposed APViT absorbs the merits of both metric learning and affinity learning methods by learning long-range spatial dependencies to model spatially-aware voxel semantics in an explicit and flexible optimization strategy.

*Corresponding author

networks (CNNs) are trained to iteratively extend one neuron object at a time, but this process is time-consuming and inappropriate for large-scale applications. The boundary detection based paradigm [Funke *et al.*, 2018; Lee *et al.*, 2017] tends to adopt the CNN in pursuit of the relationship between voxel pairs for perceiving neuron boundaries, which are post-processed to yield neuron segmentation in a gradual agglomeration manner (e.g., mutex watershed [Wolf *et al.*, 2018]). Our work follows the boundary detection based paradigm credited to achieving competitive performance (e.g., top tier in SNEMI3D [Lee *et al.*, 2017] neuron segmentation challenge) while sustaining an efficient pipeline for numerous neuron objects.

In the boundary detection based paradigm, two representative methods are affinity learning and metric learning conditioned on the network optimization strategy [Huang *et al.*, 2022]. On one hand, affinity learning methods [Funke *et al.*, 2018; Lee *et al.*, 2017; Beier *et al.*, 2017] transform the ground truth into affinity to constrain the network to learn the relationship between voxel pairs in an intuitive yet implicit optimization manner, seeking perceptibility to spatial position and discrimination to adjacent neuron instances with similar appearance (see Figure 1b). However, these methods directly output multi-channel maps as affinities, which tend to suffer from the absence of explicit voxel semantics, leading to confusion about long-range voxel correlation (affinity). On the other hand, the work [Lee *et al.*, 2021] takes advantage of metric learning to pull voxels belonging to the same neuron instance together and push those of unrelated instances away on top of manual threshold based optimization, performing well in preserving instance semantic information well, which is crucial for further improvements in accuracy (see Figure 1a). However, hand-crafted threshold is fragile in flexibility and robustness for different datasets. Moreover, relying solely on the optimization of voxel embeddings inevitably compromises the spatial information, and this negative impact is inevitably amplified by inbuilt localized receptive fields of 3D CNNs. Overall, the above analysis indicates that affinity learning methods and metric learning methods are naturally complementary. The former preserves the spatial information well and possesses a more flexible yet implicit optimization, but suffers from the absence of explicit voxel semantics, while the latter is just the other way around. Besides, the inherent locality of 3D CNNs limits both formulations to modeling long-range dependencies and capturing global voxel context, leading to sub-optimal results. Therefore, it is more desirable to integrate these two formulations to exploit their complementary potential by learning long-range spatial dependencies to model spatial-aware voxel semantics in an explicit and flexible optimization strategy.

Motivated by the above discussions, we propose a coherent and unified Appearance Prompt Vision Transformer (APViT) to enable appearance knowledge conditioned on affinity to instruct voxels with explicit semantics from a global perspective based on metric learning (Figure 1c), including an extension continuity-aware attention module and an appearance prompt modulator. **In the extension continuity-aware attention module**, we construct hierarchical attention customized for neuron extensibility and slice continuity to learn instance voxel semantic context from a global perspective and utilize continuity priors to enhance voxel spatial awareness. (a) Neuron extensibility. Considering that interleaved different neuron instances contain intricate morphology, which tends to extend from one end of the input 3D volume to another. Therefore, to endow voxel semantics with long-range dependencies, we take advantage of the extension-aware attention mechanism to aggregate global context to each voxel position to obtain robust context-aware voxel embeddings that can adapt to extended neuronal morphology. (b) Slice continuity. Intuitively, the neural deformation across several contiguous slices is always smooth and continuous. Thus, we employ the continuity-aware attention mechanism for aggregating information in a corresponding 3D spatial neighborhood for each voxel location, aiming to empower the discrimination to neighboring neuron instances. **In the appearance prompt modulator**, we draw inspiration from the prompt-based learning [Jia *et al.*, 2022], which provides a general paradigm for specific knowledge learning from offline training, and leverage voxel-adaptive appearance knowledge conditioned on affinity rich in spatial information to instruct instance voxel semantics, exploiting the potential of affinity learning to complement metric learning. In specific, we prepend a set of appearance prompts encapsulated in prompt base to interact with voxel features by cross-attention mechanism to obtain voxel-adaptive appearance prompts, achieving appearance knowledge extraction. Then the resultant voxel-adaptive appearance prompts are leveraged to modulate the voxel feature rich in semantic pattern conditioned on the calculated affinity to enhance spatial awareness. Besides, we impose the diversity loss to expand the discrepancy among prompts, enabling prompts to carry diverse and comprehensive knowledge for voxel appearance. For training, we optimize the model with centroid-anchored contrastive learning to well structure the voxel embedding space against the coarseness of manual threshold.

An attention-based task information modeling algorithm is proposed. To solve the problem that the average pooling operation in traditional task embedding generation methods is too coarse, this algorithm introduces the attention mechanism to capture the important difference of different samples, so as to extract more accurate task information. The algorithm utilizes learnable task vectors to store task information and uses an attention mechanism to identify and assign high weights to critical samples, then aggregate sample features to model task information. To verify the effectiveness of the proposed method, extensive experiments are carried out on several standard few-shot image classification datasets. Experimental results show that the proposed attention based task information modeling algorithm achieves better performance compared with the existing methods. Furthermore, our APViT can adapt to multiple post-processing operations (e.g., waterz [Funke *et al.*, 2018]), in other words, it could be possible to enjoy the flexibility with a single trained model via adaptive modulation of the post-processing configuration at the test time.

The contributions of our method could be summarized as follows: (1) We propose an appearance prompt vision transformer tailored for the connectome reconstruction in a coherent and unified framework. Specifically, we design the

extension continuity-aware attention module to construct hierarchical attention customized for neuron extensibility and slice continuity, the appearance prompt modulator to exploit the potential of affinity learning to complement metric learning. (2) To the best of our knowledge, this is the first work to absorb the merits of both affinity learning and metric learning formulation by learning long-range spatial dependencies to model spatially-aware voxel semantics in an explicit and flexible optimization strategy. (3) Extensive experimental results on multiple challenging benchmarks demonstrate that our APViT achieves consistent improvements with huge flexibility under the same post-processing strategy.

2 Related Work

2.1 Connectome Reconstruction

The reconstruction of connectomes has tremendous biological significance for studying neuronal morphology and activity. Deep learning-based methods have paved the way for research, which can be roughly divided into two categories: object tracking based methods and boundary detection based methods. Among them, the object tracking based methods [Januszewski *et al.*, 2018; Meirovitch *et al.*, 2016] only reconstruct a single neuron at a time, which is inefficient and time-consuming. In contrast, boundary detection based methods exhibit superior performance. Among them, [Funke *et al.*, 2018; Lee *et al.*, 2017] take advantage of affinity learning to separate connectomes with similar appearances. [Lee *et al.*, 2021] aggregates voxels belonging to the same connectome with the help of metric learning and utilizes handcrafted threshold to optimize the reconstruction result. However, the affinity learning paradigm tends to suffer from the absence of explicit voxel semantics, while the metrics learning one inevitably drops the spatial information, and it is greatly affected by the artificial threshold. [Huang *et al.*, 2022] absorbs the merits of both affinity learning and metric learning methods but fails to model long-range dependencies due to the locality of 3D CNNs. Different from those methods, we propose to integrate affinity learning and metric learning via a unified Appearance Prompt Vision Transformer to alleviate the above problems and accomplish the task of connectome reconstruction in an explicit yet flexible manner.

2.2 Vision Transformer and Prompt Learning

Vision Transformer. Transformer was originally introduced in [Vaswani *et al.*, 2017] for machine translation. Many efforts [Sun *et al.*, 2021; Wang *et al.*, 2022; Mai *et al.*, 2023; Luo *et al.*, 2023; Wang *et al.*, 2023; Chen and Lian, 2022] have also been made to apply it to vision tasks, including object detection, image classification and image segmentation. ViT [Dosovitskiy and Beyer, 2020] applies a transformer architecture on sequences of image patches to capture global cues for image classification tasks, building a new foundation for numerous vision tasks. Besides, emerging from transformer, prompt-based learning has been proven effective in NLP tasks by importing language instruction (prompt) to the input text so that the language model can perform well for the downstream tasks. For example, VPT [Jia *et al.*, 2022] dynamically learns a set of trainable prompts to acquire task-specific information. Nonetheless, the remaining problem is that it is not suitable to learn generic prompts for scenario adaptation. Hence we design a prompt-aware transformer to model adaptive prompts for different connectomes.

Prompt Learning. Prompting [Liu *et al.*, 2021] originally refers to inserting a few instructions to the input sentences in the NLP tasks [Gao *et al.*, 2021]. Many recent works [Li and Liang, ; Gu *et al.*, 2022] propose to exploit the prompting techniques to deal with different downstream tasks or domains with the combination of transformers without optimizing all of the parameters. In this paper, we prepend a set of appearance prompts to modulate the voxel embedding for better instructing instance voxel semantics.

3 Method

In this section, we first present the overall architecture of the Appearance Prompt Vision Transformer (APViT) as shown in Figure 2, and then describe each component in detail.

3.1 Overview

APViT enables appearance knowledge conditioned on affinity to instruct voxels from a global perspective based on metric learning, that is, exploits the complementary potential of metric learning and affinity learning, and has four stages (indexed with i). Each stage of APViT encapsulates a patch embedding, L_i extension continuity-aware attention module (ECAM, Section 3.2), and an appearance prompt modulator (APM, Section 3.3). For training, we optimize the model with centroid-anchored contrastive learning (Section 3.4) to well structure the voxel embedding space against the coarseness of manual threshold in previous work.

In specific, given the input neuron volume $I \in \mathbb{R}^{D\times H\times W}$, where D, H, and W denote depth, height and width, respectively. In the first stage, we first divide I into $D \times \frac{H}{2} \times \frac{W}{2}$ patches, each of size $1 \times 2 \times 2$. Then, we feed the flattened patches to a linear projection and obtain embedded patches of size $\frac{DHW}{2^2} \times C_1$. After that, the embedded patches are passed through a APViT block, and the output is reshaped to a feature map $\mathbf{F}_1$ of size $D \times \frac{H}{2} \times \frac{W}{2} \times C_1$. In the same way, at the beginning of each stage i, using the feature map from the previous stage as input, we obtain the feature map $\mathbf{F}_i$ of size $D \times \frac{H}{P_i} \times \frac{W}{P_i} \times C_i$, where $P_i = 2^i$, and $i = \{1, 2, 3, 4\}$.

3.2 Extension Continuity-aware Attention Module

Considering that all previous methods based on metric learning and affinity learning employ 3D convolutions, however, the inherent locality of 3D CNNs limits both formulations to modeling long-range dependencies and capturing global voxel context, leading to sub-optimal results. Therefore, we develop an extension continuity-aware attention module to construct hierarchical attention customized for neuron extensibility and slice continuity to learn instance voxel semantic context from a global perspective and utilize continuity priors to enhance voxel spatial awareness.

Extension-aware Attention. Considering that interleaved different neuron instances contain intricate morphology, which tends to extend from one end of the input 3D volume

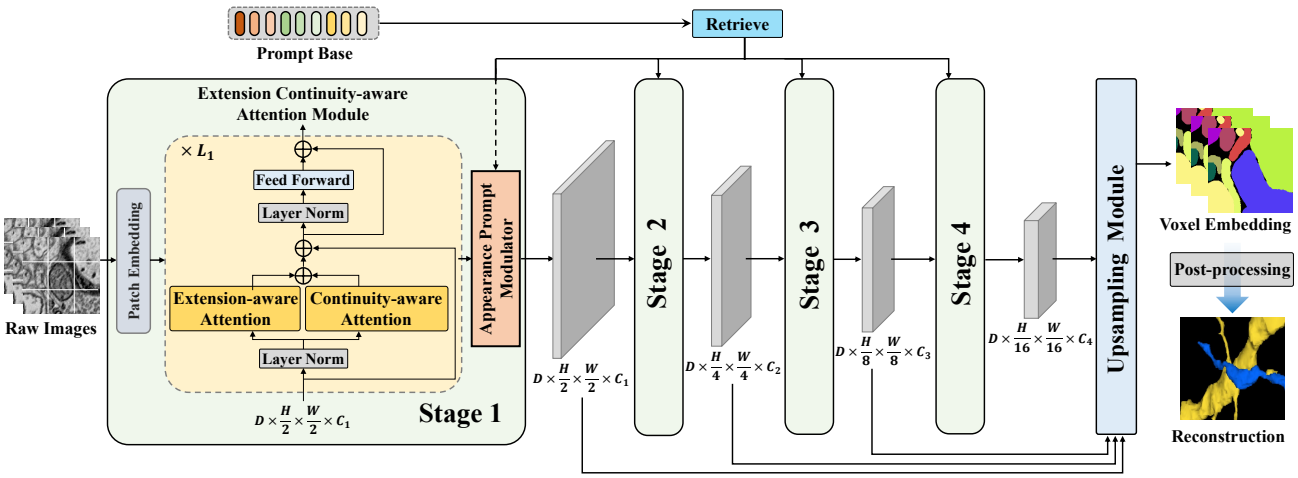

Figure 2: **The overview of the APViT framework.** Raw images are processed in four consecutive stages, each stage of APViT encapsulates a patch embedding, L_i extension continuity-aware attention module (ECAM, Section 3.2) to extract hierarchical features and an appearance prompt modulator (APM, Section 3.3) to learn voxel-adaptive appearance knowledge. After the upsampling module aggregating features from different stages, the final reconstruction result can be obtained through a post-processing step.

to another, we specially design an Extension-aware Attention Module ($\mathrm{Att_E}$) to obtain semantic-rich voxel embedding with long-range dependencies. Specifically, given the feature map $\mathbf{F} \in \mathbb{R}^{D \times \frac{H}{P} \times \frac{W}{P} \times C}$ (omit the subscript for brevity), we first flatten the spatial dimensions to produce a feature sequence $\widetilde{\mathbf{F}} = \mathbb{R}^{\frac{DHW}{P^2} \times C}$. Queries, keys and values arise from the feature sequence as follow:

$$\mathbf{Q} = \widetilde{\mathbf{F}}\mathbf{W}^Q, \mathbf{K} = \widetilde{\mathbf{F}}\mathbf{W}^K, \mathbf{V} = \widetilde{\mathbf{F}}\mathbf{W}^V, \tag{1}$$

where $\mathbf{W}^Q \in \mathbb{R}^{C \times C_k}$, $\mathbf{W}^K \in \mathbb{R}^{C \times C_k}$, $\mathbf{W}^V \in \mathbb{R}^{C \times C_v}$ are linear projections. Then we can calculate the extension-attention weight matrix $\mathbf{S} \in \mathbb{R}^{\frac{DHW}{P^2} \times \frac{DHW}{P^2}}$ with the scaled dot-product and the output are computed by $\mathbf{S}$-weighted summation on value $\mathbf{V}$:

$$\mathrm{Att_E}(\mathbf{F}) = \mathbf{SV} = \mathrm{Softmax}(\frac{\mathbf{QK}^\top}{\sqrt{C}})\mathbf{V}, \tag{2}$$

where $\sqrt{C}$ is a scaling factor for stabilizing the training and the $\top$ denotes the transpose operation. Following the standard transformer[Vaswani *et al.*, 2017], the Eq. 2 is implemented with the multi-head mechanism and the feed-forward network (FFN) is further applied to obtain the final output.

Continuity-aware Attention. Intuitively, the neural deformation across several contiguous slices is always smooth and continuous. Thus, we employ the Continuity-aware Attention mechanism ($\mathrm{Att_C}$) for aggregating information in a corresponding 3D spatial neighborhood for each voxel location, aiming to empower the discrimination to neighboring neuron instances. Predefining a spatial window $\sigma(\lambda)$ centered at voxel λ with size of $z \times p \times p$, we denotes $\mathbf{F}_{\sigma(\lambda)} \in \mathbb{R}^{zp^2 \times C}$ as the features of neighbourhood voxels and $\boldsymbol{f}_\lambda$ as the feature of voxel λ. Then, denoting $\boldsymbol{f}_\lambda$ as query, $\mathbf{F}_{\sigma(\lambda)}$ as keys and values, we can get $\mathbf{q}, \mathbf{K}, \mathbf{V}$ by:

$$\mathbf{q} = \boldsymbol{f}_\lambda \mathbf{W}^Q, \mathbf{K} = \mathbf{F}_{\sigma(\lambda)}\mathbf{W}^K, \mathbf{V} = \mathbf{F}_{\sigma(\lambda)}\mathbf{W}^V. \tag{3}$$

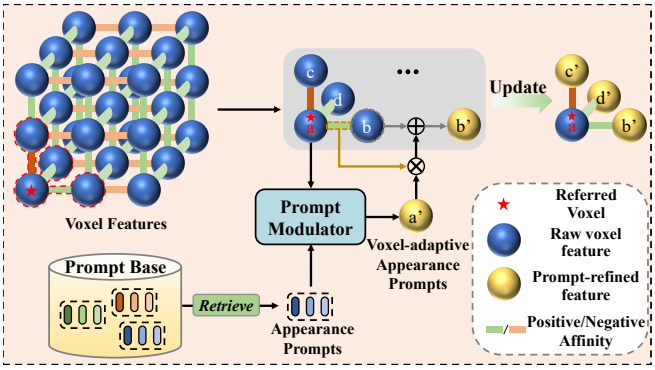

Figure 3: Illustration of the appearance prompt modulator. After calculating the affinity of each voxel in three directions, the transferred prompts are then adaptively retrieved from the prompt base to interact with the prompt modulator to update the voxel features.

Similar to Eq. 2, the output of $\mathrm{Att_C}$ can be calculated by:

$$\mathrm{Att_C}(\boldsymbol{f}_\lambda) = \mathrm{Softmax}(\frac{\mathbf{qK}^\top}{\sqrt{C}})\mathbf{V}. \tag{4}$$

Treating $D \times \frac{H}{P} \times \frac{H}{W}$ voxels for $\mathbf{F}$ in the same way, we can get $\mathrm{Att_C}(\mathbf{F})$. The above two attention layers work in parallel and the output of them are added and fed into next extension continuity-aware attention module.

3.3 Appearance Prompt Modulator

In order to exploit the potential of affinity learning to complement metric learning, we propose an appearance prompt modulator, leveraging voxel-adaptive appearance knowledge conditioned on affinity to instruct instance voxel semantics. We introduce a prompt base $\mathbf{P}^i = \{\boldsymbol{p}_n^i\}_{n=1}^{N_i}$, where N_i refers to the number of appearance prompts for stage $i \in \{1, 2, 3, 4\}$. The prompts will be leveraged to extract appearance knowledge by interaction with voxel features, then modulate the

voxel feature conditioned on the calculated affinity, which highly represents spatial association between voxels. In specific, as shown in Figure 3, for an arbitrary voxel a with feature $\boldsymbol{f}_a^i$, we calculate its affinity with adjacent voxel b, c, d at 3 directions, respectively. For simplicity, take b as example, the affinity can be formulated as

$$A_{a,b} = \max\left(0, \frac{(\boldsymbol{f}_a^i)^\top \boldsymbol{f}_b^i}{\|\boldsymbol{f}_a^i\|_2 \left\|\boldsymbol{f}_b^i\right\|_2}\right). \tag{5}$$

The affinity describes the similarity between two spatially adjacent voxels. If the affinity is close to 1, it indicates that the two voxels have a high probability of belonging to the same instance; otherwise, 0 indicates different instances. We use the voxel feature $\boldsymbol{f}_a^i$ to retrieve the appearance prompts from prompt base $\mathbf{P}^i$ and obtain voxel-adaptive prompts, as

$$\hat{\boldsymbol{p}}_a^i = \boldsymbol{f}_a^i + \mathrm{Softmax}\left(\frac{\boldsymbol{f}_a^i(\mathbf{P}^i)^\top}{\sqrt{C^i}}\right)\mathbf{P}^i. \tag{6}$$

The voxel-adaptive appearance prompt $\hat{\boldsymbol{p}}_a^i$ is the linear combination of the prompts conditioned on voxel a, and is utilizing to modulate the feature of voxel b as follow,

$$\hat{\boldsymbol{f}}_b^i = \boldsymbol{f}_b^i + A_{a,b} \cdot \hat{\boldsymbol{p}}_a^i. \tag{7}$$

In order to enable prompts to carry diverse and comprehensive knowledge for voxel appearance, we impose the diversity loss on the prompt base $\mathbf{P}$. Formally,

$$\mathcal{L}_{div} = \sum_{i=1}^{4}\sum_{m=1}^{N_i}\sum_{n=1,m\neq n}^{N_i}\cos(\mathbf{p}_m^i, \mathbf{p}_n^i), \tag{8}$$

where the $\cos(\cdot,\cdot)$ denotes cosine similarity.

3.4 Training Objectives

After consecutive four stages process, the output of each stage will be fed into an upsampling module [Hatamizadeh *et al.*, 2022] to restore the original size and get the final voxel embedding $\mathbf{E} \in \mathbb{R}^{D\times H\times W\times C}$. To well structure the voxel embedding space, we design two centroid-anchored contrastive loss. Firstly, we calculate the centroid $\overline{\mathbf{e}}_i$ for each instance i, by averaging the embedding of the voxels belonging to instance i based on the ground truth. With the set of centroid $\{\overline{\mathbf{e}}_i\}_{i=1}^N$ (N denotes the number of instances), we can get:

$$\mathcal{L}_{c1} = \sum_{i=1}^{N}\sum_{\mathbf{e}\in\mathcal{E}_i} -\log\frac{\exp(\mathbf{e}^\top\overline{\mathbf{e}}_i/\varepsilon)}{\exp(\mathbf{e}^\top\overline{\mathbf{e}}_i/\varepsilon)+\sum_{\overline{\mathbf{e}}^-\in\mathcal{E}^-}\exp(\mathbf{e}^\top\overline{\mathbf{e}}^-/\varepsilon)}, \tag{9}$$

where $\mathcal{E}_i$ denotes the set of voxel embeddings belonging to instance i, $\mathcal{E}^- = \{\overline{\mathbf{e}}_j\}_{j=1}^N / \overline{\mathbf{e}}_i$, and the temperature ε controls the concentration level of representations. Intuitively, Eq. 9 enforces each voxel embedding $\mathbf{e}$ to be similar with its ground truth ('positive') centroid and dissimilar with other irrelevant ('negative') centroids. Another contrastive loss is proposed for compactness by directly minimizing the distance between each embedded voxel and its ground truth centroid:

$$\mathcal{L}_{c2} = \sum_{\mathbf{e}\in\mathcal{E}_i}(1-\mathbf{e}^\top\overline{\mathbf{e}}_i)^2. \tag{10}$$

Note that both $\mathbf{e}$ and $\overline{\mathbf{e}}_i$ are ℓ_2-normalized. As a result, our overall training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{c1} + \mathcal{L}_{c2} + \lambda_{div}\times\mathcal{L}_{div}, \tag{11}$$

where λ_{div} is the trade-off weight.

4 Experiments

4.1 Experimental Setup

Datasets. Two commonly used neuron datasets, named CREMI [Funke *et al.*,] and AC3/AC4 [Arganda-Carreras *et al.*, 2015], are used for the evaluation of our method. CREMI dataset is divided into three sub-datasets, each consisting of two volumes of size $125 \times 1250 \times 1250$ for training and testing, respectively. We use the volume with public ground truth as the training and testing set, which consists 100 and 25 slices, respectively. The AC3/AC4 is used for SNEMI3D challenge, where size of AC3 is $256 \times 1024 \times 1024$ and AC4 consists of $100 \times 1024 \times 1024$ voxels. Following the SNEMI3D challenge, We use the top 80 slices of AC4 as training set and the rest of AC4 as validation set. And the top 100 slices of AC3 are testing set.

Implementation Details. In our APViT, the number of layers is $\{1, 2, 4, 2\}$. The volume size of the input is anisotropic (18, 160, 160), and the patch size is (1, 2, 2) at each stage. During training, our model is trained with batch size of 2, using the Adam optimizer with an initial learning rate of 0.0001 for 200,000 iterations. And we constrain the output at different resolutions for each stage with GT as an auxiliary loss where we set $\lambda_{div} = 0.1$.

Evaluation Metrics. Following the conventions, VOI (variation of information) is deemed as the main evaluation metric. We also report $ARAND$ (adapted Rand error) to assess the reconstruction results. Smaller values of these two metrics indicate better segmentation performance.

4.2 Comparison with State-of-the-art Methods

Tables 1 and 2 report the neuron reconstruction performance comparison of our method and several state-of-the-art methods on AC3/AC4 and CREMI datasets, respectively. As shown in Table 1, we replace the feature extractors of several methods by our APViT. The insertion of our APViT can bring significant performance improvements to different baseline methods. For example, when we adopt MALA [Funke *et al.*, 2018] as the baseline method and LMC as the post-processing approach, the VOI and the $ARAND$ metrics improve from 1.3857 to 1.0235 and from 0.1143 to 0.0898, respectively. From Table 2, we can observe that our proposed APViT outperforms all previous methods by a substantial margin. Specifically, APViT surpasses the second-best method (PEA) on the VOI metric by 12.01%, 13.56%, 11.81% on Cremi-A, Cremi-B, Cremi-C, respectively.

4.3 Ablation Study and Analysis

To look deeper into our method, we perform a series of ablation studies on AC3/AC4 dataset with waterz as post-processing to validate the effectiveness of APViT, including the extension continuity-aware attention module (ECAM),

Method	Waterz				LMC				MWS			
	VOI_{split}	VOI_{merge}	VOI	$ARAND$	VOI_{split}	VOI_{merge}	VOI	$ARAND$	VOI_{split}	VOI_{merge}	VOI	$ARAND$
ML-De	-	-	-	-	-	-	-	-	1.5752	0.6151	2.1903	0.1964
SuperHuman	1.0910	0.3418	1.4328	0.1685	1.1443	0.2630	1.4073	0.1221	-	-	-	-
Ours	**0.9222**	**0.3305**	**1.2527**	**0.1228**	**0.9001**	**0.2208**	**1.1209**	**0.0942**	**0.7257**	**0.5017**	**1.2274**	**0.1364**
MALA	1.0988	0.2446	1.3434	0.1089	1.1457	0.2400	1.3857	0.1143	-	-	-	-
Ours	**0.8358**	**0.1945**	**1.0303**	**0.0840**	**0.8259**	**0.1976**	**1.0235**	**0.0898**	**0.8417**	**0.3399**	**1.1815**	**0.0912**
PEA	0.9116	0.2934	1.2050	0.1212	0.8999	0.2506	1.1505	0.1069	0.8522	**0.2322**	1.0844	0.0938
Ours	**0.7671**	**0.2093**	**0.9764**	**0.0775**	**0.8231**	**0.2054**	**1.0285**	**0.0940**	**0.5943**	0.3842	**0.9785**	**0.0865**

Table 1: Comparisons of different methods on AC3/AC4 dataset.

Method	Post-processing	Cremi-A				Cremi-B				Cremi-C			
		VOI_{split}	VOI_{merge}	VOI	$ARAND$	VOI_{split}	VOI_{merge}	VOI	$ARAND$	VOI_{split}	VOI_{merge}	VOI	$ARAND$
SuperHuman	Waterz	1.0581	0.3884	1.4465	0.2167	0.8095	**0.1469**	0.9564	0.0443	0.9791	0.3992	1.3782	0.1563
	LMC	1.0883	0.4232	1.5114	0.2438	0.8281	0.1867	1.0148	0.0468	1.0017	0.2742	1.2760	0.1202
MALA	Waterz	0.5508	**0.2371**	0.7879	0.1251	0.8810	0.1685	1.0496	0.0482	1.1493	**0.1963**	1.3456	0.1308
	LMC	0.5263	0.2596	0.7859	0.1177	0.9688	0.2005	1.1694	0.0612	1.2016	0.2371	1.4387	0.1365
PEA	Waterz	0.4892	0.3001	0.7892	0.1546	0.6887	0.1978	0.8865	0.0370	1.0247	0.2255	1.2502	0.1128
	LMC	0.4774	0.2917	0.7691	0.1425	0.6648	0.2183	0.8831	0.0393	0.9983	0.2490	1.2472	0.1146
Ours	Waterz	0.4447	0.2595	**0.7041**	**0.1169**	**0.5793**	0.2014	**0.7807**	**0.0319**	**0.8839**	0.2341	**1.1181**	**0.1102**
	LMC	**0.4336**	0.2914	0.7249	0.1304	0.5777	0.2162	0.7939	0.0340	0.8719	0.2527	1.1247	0.1116

Table 2: Comparisons of different methods on CREMI dataset.

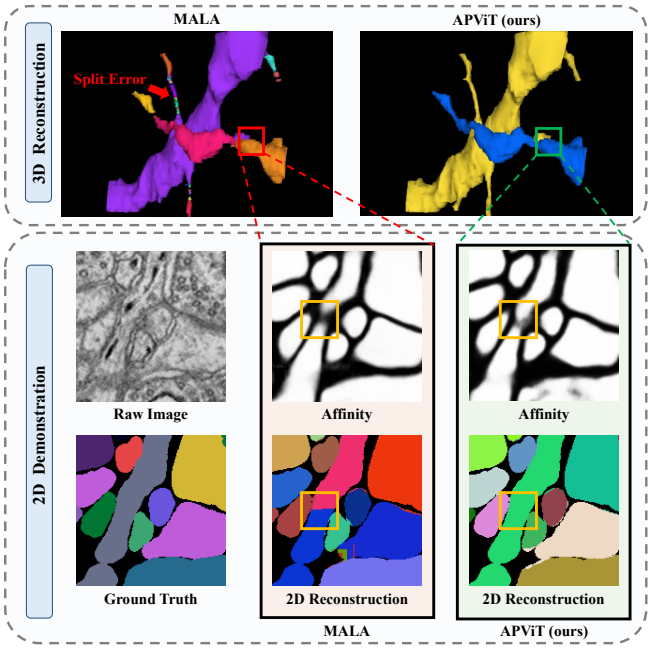

Figure 4: Comparison between MALA and APViT (ours) on AC3/AC4 datatset.

Model	#Params	VOI	$ARAND$
CNN-based model	36.78M	1.1989	0.1685
Transformer-based model	35.43M	1.4073	0.1321
Ours	37.25M	**0.9764**	**0.0775**

Table 3: Illustration of advantages of APViT on AC3/AC4 dataset.

ECAM	APM	Contrastive learning	VOI	$ARAND$
			1.1989	0.1932
✓			1.0998	0.1095
	✓		1.1698	0.1313
✓	✓		1.0282	0.0925
✓	✓	✓	**0.9764**	**0.0775**

Table 4: Ablation on main components on AC3/AC4 dataset.

the appearance prompt modulator (APM), and the centroid-anchored contrastive learning. Note that we remove all proposed modules and only maintain the bald vision transformer, and take manual threshold based optimization following [Lee *et al.*, 2021] as our baseline.

Effectiveness of Main Components. Table 4 summarizes the results of module ablation studies under different configurations. (1) We ablate the ECAM to study the importance of hierarchical attention. As deteriorated results indicate, customized for neuron extensibility and slice continuity is crucial to learn voxel semantic context from a global perspective and utilize continuity priors to enhance voxel spatial awareness. (2) Then we investigate the impact of introducing APM, and observe a absolute performance lift (from 0.1932 to 0.1313 in $ARAND$). The improvements can be mainly ascribed to the strong ability of the APM to leverage voxel-adaptive appearance knowledge conditioned on affinity to instruct voxel semantics, exploiting the potential of affinity learning to complement metric learning. (3) We also explore the centroid-anchored contrastive learning. When we replace the optimization strategy with manual threshold based optimization [Lee *et al.*, 2021], the performance of the model is clearly degraded. This proves the necessity of contrastive

Extension-aware attention	Continuity-aware attention	VOI	$ARAND$
		1.1698	0.1313
✓		1.1310	0.1083
	✓	1.0560	0.0899
✓	✓	**0.9764**	**0.0775**

Table 5: Ablation on different attention mechanisms on AC3/AC4.

Window size at Stage 4 ($z \times x \times y$)	*VOI*	*ARAND*
$1 \times 7 \times 7$	1.1084	0.1415
$3 \times 5 \times 5$	1.0285	0.0845
$5 \times 3 \times 3$	1.0169	0.0795
$7 \times 1 \times 1$	**0.9764**	**0.0775**

Table 6: Ablation on window sizes in continuity-aware attention.

Numbers of prompt at each stage (from 1 to 4)	*VOI*	*ARAND*
(3, 3, 3, 3)	1.0125	0.0841
(6, 6, 6, 6)	1.0804	0.0956
(9, 9, 9, 9)	1.0544	0.0932
(12, 12, 12, 12)	1.1329	0.1510
(12, 9, 6, 3)	**0.9764**	**0.0775**

Table 7: Ablation on different prompt numbers at each stage.

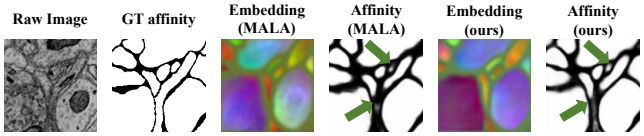

Figure 5: Visualization of the affinity and embedding.

learning to well structure the voxel embedding space against the coarseness and sensitivity of manual threshold. Without all the proposed methods, the model has degenerated into the baseline. The performance improvement of our final model over the baselines is significant.

Advantages of Our Framework. To validate the advantage of our framework tailored for connectome reconstruction, we perform an ablation study to investigate the impact of 3D CNNs (3D ResUNet), pure vision transformer (UNETR [Hatamizadeh *et al.*, 2022]), and our APViT with the same parameters, as tabulated in Table 3. We observe that transformer-based method outperforms CNN-based methods due to long-range dependency modeling capabilities. Furthermore, our APViT achieves a significant lead, which indicates that instead of simply using the vision transformer, APViT absorbs the merits of both affinity learning and metric learning formulation to model spatially-aware voxel semantics in an explicit and flexible optimization strategy. More importantly, it could be possible to enjoy the flexibility with a single trained model via adaptive modulation of the post-processing configuration at the test time.

Effectiveness of Extension Contiguity-aware Attention. To analyze the ECAM in depth, we ablate extension-aware attention and continuity-aware attention separately, as described in Table 5. Adding either of the two attention mechanisms contributes to a remarkable performance gain. Furthermore, these two mechanisms work in conjunction enables further performance gain, benefiting from learning instance voxel semantic context from a global perspective and utilizing continuity priors to enhance voxel spatial awareness.

Local Window Size. In Table 6, we observe that the local window size in continuity-aware attention has a large impact on reconstruction performance. Experiments show that APViT achieves the best result at the window size of $7 \times 1 \times 1$. We conjecture that voxels in the z-stereoscopic direction contain more connectome information, including connections between slices, thus larger size along the z-axis of the local window is more beneficial for feature extraction.

Prompt Strategy. As shown in Table 7, we ablate the components inside the prompt assignment module. In fact, the introduction of a prompt can be deemed as a guidepost in artificially solving incidents, i.e., guiding information, which can strongly correct voxel features based on ECAM and enhance different instance clustering to avoid foreground and background confusion issues better, thus reducing splitting and fusion errors. Therefore, the number of prompts selected also plays an important role in the prompt assignment module, directly affecting the modeling ability of prompts. Compared with using a fixed number of prompts at each stage, the flexible, prompt allocation strategy is obviously more advantageous for the diverse neuron reconstruction task, yielding significant improvements. The principal reason for this phenomenon is that the feature resolution processed by each encoder stage is different, making it necessary to customize the number of prompts according to different stages.

4.4 Explainable Visualization Study

We visualize the reconstruction results of various methods on the test set of AC3/AC4. It can be seen that each method performs well when reconstructing relatively simple and clear neurons, however, when faced with entangled adjacent neuron individuals, the previous methods generally perform unsatisfactorily. In Figure 4, MALA produces more over-segmentation results (e.g., blue and red connectomes), separating neurons that should belong to the same label, resulting in split errors, while APViT is able to reconstruct the correct neuron. The underlying reason is that the extension continuity-aware attention module takes into account a wide range of each neuron. Meanwhile, the continuity-aware attention mechanism takes advantage of the smoothness and continuity between slices, leading to less split error in ambiguous areas. In Figure 5, it can be observed that the embedding is the clear cluster corresponding to each neuron instance. And the affinity map obtained from our voxel embeddings does not misjudge the confusing boundaries. It indicates that the appearance prompt modulator aggregates rich spatial information via the voxel-adaptive appearance prompt.

5 Conclusion

In this paper, we propose we propose a coherent and unified Appearance Prompt Vision Transformer (APViT) to enable appearance knowledge conditioned on affinity to instruct voxels with explicit semantics based on metric learning, including an extension continuity-aware attention module and an appearance prompt modulator. Extensive experimental results on challenging benchmarks show effectiveness.

Contribution Statement

Rui Sun, Naisong Luo and Yuwen Pan contributed equally to this paper.

Acknowledgments

This work was partially supported by the National Nature Science Foundation of China (Grant 62022078, 62021001).

References

[Arganda-Carreras *et al.*, 2015] Ignacio Arganda-Carreras, Srinivas C Turaga, and Berger. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in neuroanatomy*, page 142, 2015.

[Ascoli, 2002] Giorgio A Ascoli. *Computational neuroanatomy: Principles and methods*. Springer Science & Business Media, 2002.

[Beier *et al.*, 2017] Thorsten Beier, Constantin Pape, Nasim Rahaman, Timo Prange, Stuart Berg, Davi D Bock, Albert Cardona, Graham W Knott, Stephen M Plaza, Louis K Scheffer, et al. Multicut brings automated neurite segmentation closer to human performance. *Nature methods*, 14(2):101–102, 2017.

[Berning *et al.*, 2015] Manuel Berning, Kevin M Boergens, and Moritz Helmstaedter. Segem: efficient image analysis for high-resolution connectomics. *Neuron*, 87(6):1193–1206, 2015.

[Chen and Lian, 2022] Zenggui Chen and Zhouhui Lian. Semi-supervised semantic segmentation via prototypical contrastive learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6696–6705, 2022.

[Donohue and Ascoli, 2011] Duncan E Donohue and Giorgio A Ascoli. Automated reconstruction of neuronal morphology: an overview. *Brain research reviews*, 67(1-2):94–102, 2011.

[Dosovitskiy and Beyer, 2020] Alexey Dosovitskiy and Beyer. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Funke *et al.*,] Jan Funke, Eric Perlman, Srini Turaga, Davi Bock, and Stephan Saalfeld. Cremi challenge leaderboard, as of 2017/22/09.

[Funke *et al.*, 2018] Jan Funke, Fabian Tschopp, William Grisaitis, Arlo Sheridan, Chandan Singh, Stephan Saalfeld, and Srinivas C Turaga. Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *TPAMI*, 41(7):1669–1680, 2018.

[Gao *et al.*, 2021] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *ACL*, pages 3816–3830. ACL, 2021.

[Gu *et al.*, 2022] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. PPT: pre-trained prompt tuning for few-shot learning. In *ACL*, pages 8410–8423. Association for Computational Linguistics, 2022.

[Hatamizadeh *et al.*, 2022] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[Huang *et al.*, 2022] Wei Huang, Shiyu Deng, Chang Chen, Xueyang Fu, and Zhiwei Xiong. Learning to model pixel-embedded affinity for homogeneous instance segmentation. In *AAAI*, 2022.

[Januszewski *et al.*, 2018] Michał Januszewski, Jörgen Kornfeld, Peter H Li, Art Pope, Tim Blakely, Larry Lindsey, Jeremy Maitin-Shepard, Mike Tyka, Winfried Denk, and Viren Jain. High-precision automated reconstruction of neurons with flood-filling networks. *Nature methods*, 15(8):605–610, 2018.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

[Lee *et al.*, 2017] Kisuk Lee, Jonathan Zung, Peter Li, Viren Jain, and H Sebastian Seung. Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*, 2017.

[Lee *et al.*, 2021] Kisuk Lee, Ran Lu, Kyle Luther, and H Sebastian Seung. Learning and segmenting dense voxel embeddings for 3d neuron reconstruction. *TMI*, 40(12):3801–3811, 2021.

[Li and Liang,] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *ACL*.

[Liu *et al.*, 2021] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586, 2021.

[Livet and Weissman, 2007] Jean Livet and Tamily Weissman. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450(7166):56–62, 2007.

[Luo *et al.*, 2023] Naisong Luo, Yuwen Pan, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Camouflaged instance segmentation via explicit de-camouflaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[Mai *et al.*, 2023] Huayu Mai, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Dualrel: semi-supervised mitochondria segmentation from a prototype perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[Meirovitch *et al.*, 2016] Yaron Meirovitch, Alexander Matveev, Hayk Saribekyan, David Budden, David Rolnick, Gergely Odor, Seymour Knowles-Barley, Thouis Raymond Jones, Hanspeter Pfister, Jeff William Lichtman, et al. A multi-pass approach to large-scale connectomics. *arXiv preprint arXiv:1612.02120*, 2016.

[Meirovitch *et al.*, 2019a] Yaron Meirovitch, Lu Mi, Hayk Saribekyan, Alexander Matveev, David Rolnick, and Nir Shavit. Cross-classification clustering: an efficient multi-object tracking technique for 3-d instance segmentation in connectomics. In *CVPR*, pages 8425–8435, 2019.

[Meirovitch *et al.*, 2019b] Yaron Meirovitch, Lu Mi, Hayk Saribekyan, Alexander Matveev, David Rolnick, and Nir Shavit. Cross-classification clustering: an efficient multi-object tracking technique for 3-d instance segmentation in connectomics. In *CVPR*, pages 8425–8435, 2019.

[Sun *et al.*, 2021] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 30, 2017.

[Wang *et al.*, 2022] Yuan Wang, Rui Sun, Zhe Zhang, and Tianzhu Zhang. Adaptive agent transformer for few-shot segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 36–52. Springer, 2022.

[Wang *et al.*, 2023] Yuan Wang, Rui Sun, and Tianzhu Zhang. Rethinking the correlation in few-shot segmentation: a buoys view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[Wolf *et al.*, 2018] Steffen Wolf, Constantin Pape, Alberto Bailoni, Nasim Rahaman, Anna Kreshuk, Ullrich Kothe, and FredA Hamprecht. The mutex watershed: efficient, parameter-free image partitioning. In *ECCV*, pages 546–562, 2018.