# TopoTTA: Topology-Enhanced Test-Time Adaptation for Tubular Structure Segmentation

Supplementary Material

## A. Overview

This appendix is structured as follows:

- In Sec. B, we provide more information on implementation details, including the process of TopoTTA, the implementation details in Stage 2, resizing the ground-truth label, more TopoMDCs, and training the source model.
- In Sec. C, we present additional experiment results, *i.e.* additional comparison results, additional ablation study, and additional visualization results.[1]

## B. More Implementation Details

### B.1. The Process of TopoTTA

The overall process of TopoTTA is summarized in Algorithm 1.

**Algorithm 1:** The Process of TopoTTA

**Input :** A source pre-trained TSS model $\mathcal{F}(\cdot;\theta)$, teacher model $\mathcal{F}(\cdot;\theta')$ target domain dataset $\mathcal{D}^t = \{\boldsymbol{x}_i^t\}_{i=1}^{N^t}$, learning rates $\alpha_1$ and $\alpha_2$, number of iteration $iterations$

**Output:** Final prediction $\hat{\boldsymbol{y}}_{\text{out}}$

**for** $\boldsymbol{x} \in \mathcal{D}^t$ **do**
  # Stage 1: Topological structure adaptation
  Define TopoMDCs by Eq. (3,4,5);
  $\mathcal{E} \leftarrow$ Encoder $(\mathcal{F}(\cdot;\theta))$; # Extract encoder from $\mathcal{F}(\cdot;\theta)$
  **for** $3 \times 3$ $Conv$ *in* $\mathcal{E}$ **do**
    Replace $Conv$ with TopoMDCs$(\cdot;\theta;\boldsymbol{\delta})$;
  **end**
  **for** $j \leftarrow 1$ ***to*** $iterations/2$ **do**
    $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} - \alpha_1 \cdot \nabla_{\boldsymbol{\delta}} \mathcal{L}_{\text{EM}}(\mathcal{F}(\boldsymbol{x};\theta;\boldsymbol{\delta}))$;
  **end**
  # Stage 2: Topological continuity refinement
  **for** $j \leftarrow 1$ ***to*** $iterations/2$ **do**
    $\hat{\boldsymbol{y}}' \leftarrow \mathcal{F}(\boldsymbol{x};\theta';\boldsymbol{\delta})$;
    Select $N_p$ points as key points;
    **for** $p = (u_c, v_c) \leftarrow 1$ ***to*** $N_p$ **do**
      $\boldsymbol{x}_p^{\text{fg}}$ centered at point $p$, with a size of $s \times s$ ;
      $\boldsymbol{x}_p^{\text{bg},*} \leftarrow \arg\min_{\boldsymbol{x}_p^{\text{bg}}} Confidence(\hat{\boldsymbol{y}}_p'^{;\text{bg}})$; # $\boldsymbol{x}_p^{\text{bg}}$ denotes the background sliding window around $\boldsymbol{x}_p^{\text{fg}}$
      Obtain $\boldsymbol{x}_p^{\text{swap}}$ using low-frequency swapping by Eq. (7);
      Obtain *pseudo-break* patch $\boldsymbol{x}_p'$ by Eq. (8);
    **end**
    $\theta \leftarrow \theta - \alpha_2 \cdot \nabla_{\theta} \mathcal{L}_{\text{CE}}(\hat{\boldsymbol{y}}', \mathcal{F}(\boldsymbol{x}';\theta;\boldsymbol{\delta}))$;
  **end**
  # Prediction
  $\hat{\boldsymbol{y}}_{\text{out}} \leftarrow \mathcal{F}(\boldsymbol{x};\theta;\boldsymbol{\delta})$
**end**

[1] Code will be released at https://anonymous.4open.science/r/TopoTTA-82A0.

## B.2. Other Implementation Details in Stage 2

For the Stage 2, we follow the CoTTA's settings[60]. Before inputting the image $\boldsymbol{x}$ into the teacher model, it undergoes four rounds of augmentations, combining random horizontal flips, vertical flips, and scaling by factors of 0.5, 1.0, 1.25, and 1.5. The average prediction from these augmented images is used as the pseudo-label. In the student model, the hard sample $\boldsymbol{x}'$ undergoes a single round of the same augmentation combinations. The Baseline$^\star$ in the ablation study already incorporates these settings.

## B.3. Resizing the Ground-Truth Label

In the manuscript, we discuss resizing the images in datasets from three scenarios to $384 \times 384$. However, conventional nearest or linear interpolation methods often cause breakage in the thin annotated regions of the ground-truth labels, negatively affecting both training and topological metric calculations. To address this issue, we propose a novel resizing method. Specifically, the images are first resized using the area-based interpolation method (available in the OpenCV library). Binarization is then applied with thresholds of 0.5 and 0, as shown in the third and fourth columns of Fig. B.1. Next, the skeleton map is extracted from the binarized result using a threshold of 0. The skeleton map and the binarized image obtained with a threshold of 0.5 are combined using a pixel-wise OR operation to produce the final resized ground-truth labels, as shown in the fifth column of Fig. B.1. This resizing method produces results that closely resemble the original ground-truth labels and effectively preserves topological consistency.

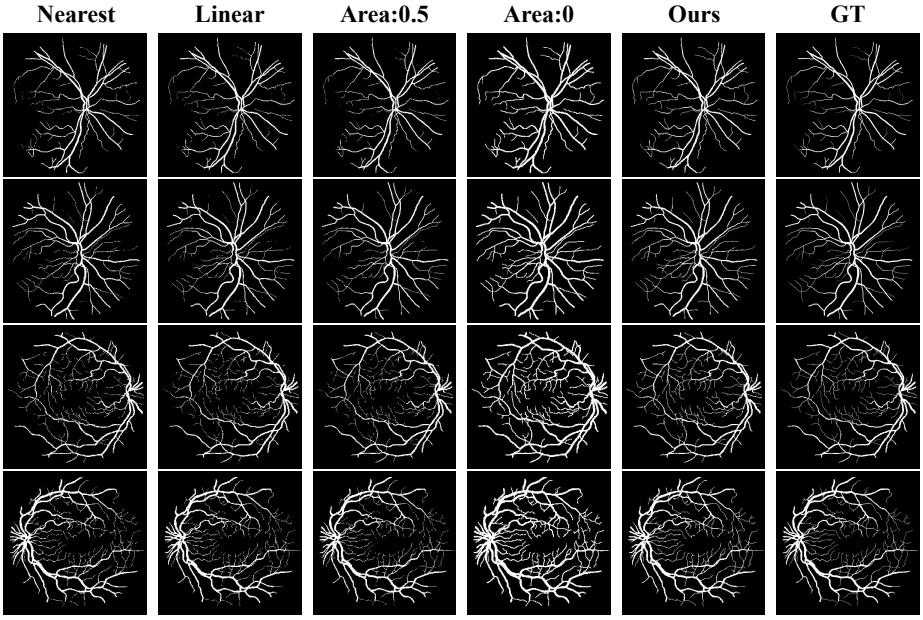


Figure B.1. Visualization of the proposed resizing method compared with other commonly used approaches. Area:0.5 represents the image binarized with a threshold of 0.5, Area:0 uses a threshold of 0, and GT is short for ground-truth labels.

## B.4. More Topology-Meta Difference Convolutions

In the manuscript, we present the formulation for calculating $\mathcal{C}_1$. Here, we provide the formulations for all other directions of TopoMDCs,

$$\begin{aligned}\mathcal{C}_i(r_x, r_y) = \mathcal{C}_c(r_x, r_y) &- \sum_{(\Delta r_x, \Delta r_y) \in \mathcal{R}_i} w(\Delta r_x, \Delta r_y) \cdot \boldsymbol{x}_{\text{in}}(r_x, r_y) \\ &+ \sum_{(\Delta r_x, \Delta r_y) \in \mathcal{R}_i,\ (\Delta b_x, \Delta b_y) \in \mathcal{B}_i} w(\Delta r_x, \Delta r_y) \cdot \boldsymbol{x}_{\text{in}}(r_x - \Delta b_x, r_y - \Delta b_y),\end{aligned} \tag{11}$$

$$\begin{aligned}
&\mathcal{R}_1 = \{(-1,-1),(-1,0),(0,-1)\}, \quad \mathcal{B}_1 = \{(-1,-1)\}; \quad \mathcal{R}_2 = \{(0,-1),(-1,-1),(1,-1)\}, \quad \mathcal{B}_2 = \{(0,-1)\}; \\
&\mathcal{R}_3 = \{(1,-1),(0,-1),(1,0)\}, \quad \mathcal{B}_3 = \{(1,-1)\}; \quad \mathcal{R}_4 = \{(-1,0),(-1,-1),(-1,1)\}, \quad \mathcal{B}_4 = \{(-1,0)\}; \\
&\mathcal{R}_5 = \{(1,0),(1,1),(1,-1)\}, \quad \mathcal{B}_5 = \{(1,0)\}; \quad \mathcal{R}_6 = \{(-1,1),(-1,0),(0,1)\}, \quad \mathcal{B}_6 = \{(-1,1)\}; \\
&\mathcal{R}_7 = \{(0,1),(1,1),(-1,1)\}, \quad \mathcal{B}_7 = \{(0,1)\}; \quad \mathcal{R}_8 = \{(1,1),(0,1),(1,0)\}, \quad \mathcal{B}_8 = \{(1,1)\}.
\end{aligned} \tag{12}$$

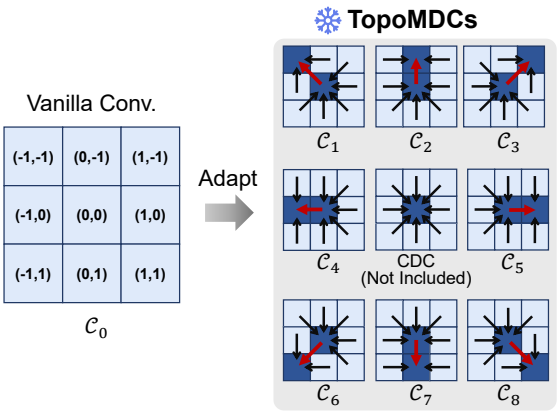


Figure B.2. Different Topology-Meta Difference Convolutions, which inherit parameters from the source domain model without adding additional parameters to convolution layers.

### B.5. Training the Source Model

The Adam optimizer with a learning rate of $5 \times 10^{-4}$ is used across all four scenarios, with the batch size set to 4. The maximum number of epochs is set to 100, 50, 100, and 60 for the four scenarios, respectively. The model with the best performance on the source domain test set is selected for testing. Training is conducted using a combination of Dice and binary cross entropy (BCE) loss functions. During training, random horizontal and vertical flips are applied as data augmentations.

## C. Additional Experimental Results

### C.1. Additional Comparison Results

Table C.1 presents the detailed experimental results of TopoTTA using CS2Net as the baseline network across four scenarios. The conclusions are consistent with those drawn when UNet is used as the baseline network: TopoTTA delivers significant improvements in both segmentation performance and topological connectivity in most scenarios. Table C.2 shows the performance of a TopoTTA variant applied to DSCNet. In this variant, Stage 1 is omitted, and only Stage 2 is used for parameter updates. This adjustment is necessary because DSCNet already incorporates deformable convolutional kernels, which limit the compatibility of TopoMDCs. As shown in Table C.2, even with this simplified version, TopoTTA still outperforms most comparison methods across the majority of scenarios. These additional experiments further validate the broad applicability and effectiveness of TopoTTA, demonstrating its ability to efficiently adapt to different CNN-based models. Table C.3 presents the paired t-test results of clDice. The results indicate that the $p$-value $<0.05$ almost across all datasets, inferring TopoTTA's improvement is significant.

Table C.1. Cross-domain testing results obtained using CS2Net as baseline network in four different scenarios, i.e., retinal vessel segmentation, road extraction, microscopic neuronal segmentation, and retinal OCT-angiography vessel segmentation. Source Only: Trained on the source, and tested on the target domain directly. The best and second-best results in each column are highlighted in **bold** and underline.

| Method | DRIVE → CHASE | | | DRIVE → STARE | | | CHASE → DRIVE | | | CHASE → STARE | | | DeepGlobe → MR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ |
| Source Only | 22.58 | 19.03 | 43.00 | 37.27 | 33.94 | 112.35 | 61.78 | 62.33 | 95.35 | 48.46 | 46.73 | 107.95 | 45.70 | 55.85 | 79.65 |
| TENT [58] | 64.01 | 65.22 | 33.50 | 60.93 | 55.65 | 107.80 | 67.53 | 67.14 | 88.65 | 60.88 | 56.91 | 107.75 | 42.41 | 52.49 | 77.88 |
| CoTTA [60] | 67.04 | 69.08 | 28.50 | 61.40 | 55.42 | 102.10 | 67.95 | 67.44 | 83.90 | 62.08 | 58.00 | 102.35 | **48.88** | 58.55 | 74.65 |
| SAR [43] | 63.72 | 64.86 | 33.00 | 60.77 | 55.52 | 109.15 | 67.17 | 66.83 | 88.95 | 60.81 | 56.86 | 107.10 | 43.69 | 55.02 | 77.11 |
| DIGA [61] | 63.87 | 64.49 | 34.13 | 59.47 | 54.34 | 108.40 | 67.78 | 67.34 | 87.15 | 61.67 | 57.45 | 107.70 | 43.71 | 55.02 | 76.96 |
| DomainAdaptor [72] | 58.39 | 57.34 | 41.38 | 54.76 | 49.29 | 111.10 | 65.21 | 65.30 | 89.80 | 57.11 | 54.00 | 112.40 | 43.55 | 56.09 | 77.37 |
| MedBN [44] | 58.10 | 58.38 | 34.63 | 57.40 | 51.92 | 116.40 | 65.21 | 63.01 | 90.95 | 61.76 | 56.85 | 110.35 | 43.42 | 53.49 | 77.53 |
| VPTTA [3] | 62.14 | 62.88 | 35.26 | 60.59 | 54.93 | 108.00 | 66.97 | 66.66 | 88.95 | 60.30 | 56.46 | 107.50 | 43.89 | 55.31 | 77.41 |
| **TopoTTA (Ours)** | **68.27** | **72.99** | **21.88** | **62.00** | **57.46** | **91.60** | **71.12** | **69.74** | **77.50** | **65.09** | **59.91** | **96.20** | 48.47 | **63.09** | **73.55** |

| Method | DeepGlobe → CNDS | | | Neub1 → Neub2 | | | Neub2 → Neub1 | | | ROSE → OCTA500 | | | OCTA500 → ROSE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ |
| Source Only | 84.28 | 91.57 | 10.95 | 22.00 | / | 7.56 | 61.76 | 72.72 | 7.31 | 46.94 | 53.88 | 61.00 | 72.36 | 75.60 | 18.33 |
| TENT [58] | 76.55 | 89.29 | 7.50 | 62.26 | 70.52 | 7.95 | 62.64 | 74.01 | 7.03 | 66.76 | 72.10 | 64.44 | 73.58 | 77.60 | 16.23 |
| CoTTA [60] | 75.43 | 89.72 | 6.98 | 63.88 | 72.98 | 8.22 | 63.66 | 75.87 | 5.78 | **68.82** | 75.01 | 52.48 | 72.07 | 76.34 | 16.89 |
| SAR [43] | 77.02 | 89.23 | 7.77 | 61.55 | 69.81 | 8.33 | 61.86 | 73.17 | 7.22 | 65.99 | 72.10 | 63.76 | 73.67 | 77.68 | 16.00 |
| DIGA [61] | 79.72 | 89.70 | 7.73 | 61.31 | 70.40 | 8.05 | 63.57 | 73.90 | 7.31 | 66.99 | 73.02 | 65.60 | 71.09 | 76.38 | 16.66 |
| DomainAdaptor [72] | 78.81 | 90.94 | 7.61 | 53.99 | 63.87 | 10.67 | 63.12 | 74.19 | 6.97 | 63.12 | 69.60 | 72.90 | 71.45 | 75.88 | 16.56 |
| MedBN [44] | 81.22 | 92.16 | 8.08 | 56.08 | 63.72 | 11.50 | 63.56 | 72.80 | 9.84 | 63.97 | 69.11 | 78.92 | 75.55 | 78.28 | **8.11** |
| VPTTA [3] | 77.32 | 89.52 | 7.68 | 60.87 | 69.56 | 9.28 | 62.21 | 73.47 | 7.32 | 65.32 | 71.54 | 65.28 | 73.34 | 77.41 | 15.77 |
| **TopoTTA (Ours)** | **87.89** | **96.05** | **4.24** | **66.73** | **75.81** | **6.72** | **64.00** | **76.60** | **5.31** | 67.62 | **76.65** | **45.92** | **75.57** | **78.72** | 15.67 |

Table C.2. Cross-domain testing results obtained using DSCNet as baseline network in four different scenarios, i.e., retinal vessel segmentation, road extraction, microscopic neuronal segmentation, and retinal OCT-angiography vessel segmentation. Source Only: Trained on the source, and tested on the target domain directly. The best and second-best results in each column are highlighted in **bold** and underline.

| Method | DRIVE → CHASE | | | DRIVE → STARE | | | CHASE → DRIVE | | | CHASE → STARE | | | DeepGlobe → MR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ |
| Source Only | 23.16 | 18.98 | 37.25 | 42.46 | 36.89 | 102.35 | 58.58 | 53.95 | 94.60 | 44.19 | / | 106.90 | 42.71 | 51.87 | 81.28 |
| TENT [58] | 64.17 | 65.42 | 22.38 | 56.87 | 50.50 | 105.90 | 66.75 | 62.32 | 83.80 | 62.18 | 55.28 | 102.85 | 40.30 | 47.69 | 79.94 |
| CoTTA [60] | 67.14 | 69.83 | 25.13 | 57.69 | 51.30 | 103.50 | 68.80 | 65.04 | 79.85 | 63.96 | 57.54 | 99.70 | 45.13 | 54.00 | **77.68** |
| SAR [43] | 64.08 | 65.30 | 23.25 | 56.98 | 50.67 | 105.75 | 66.57 | 62.12 | 84.30 | 62.18 | 55.32 | 102.90 | 42.61 | 51.79 | 80.48 |
| DIGA [61] | 63.87 | 66.58 | 24.38 | 57.82 | 51.76 | 105.75 | 65.58 | 60.95 | 85.15 | 61.43 | 54.50 | 104.35 | 42.83 | 54.67 | 78.63 |
| DomainAdaptor [72] | 60.42 | 43.22 | 105.90 | 49.74 | 43.22 | 105.90 | 65.60 | 61.44 | 85.25 | 61.99 | 54.02 | 103.50 | 43.14 | 53.12 | 79.24 |
| MedBN [44] | 58.36 | 58.80 | 29.63 | 54.57 | 49.68 | 96.30 | 66.15 | 64.52 | 81.55 | 61.50 | 57.72 | 102.50 | 41.71 | 49.59 | 83.96 |
| VPTTA [3] | 63.22 | 64.24 | 21.13 | 56.44 | 49.85 | 104.50 | 66.50 | 62.11 | 83.55 | 62.29 | 55.18 | 103.35 | 42.41 | 51.15 | 79.77 |
| **TopoTTA (Stage2 only)** | **67.32** | **71.89** | **19.13** | **60.24** | **53.06** | **94.25** | **70.35** | **66.88** | **79.45** | **64.63** | **58.05** | **95.35** | **47.17** | **61.55** | 77.82 |

| Method | DeepGlobe → CNDS | | | Neub1 → Neub2 | | | Neub2 → Neub1 | | | ROSE → OCTA500 | | | OCTA500 → ROSE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ |
| Source Only | 81.80 | 88.67 | 14.73 | 3.94 | / | 9.44 | 58.66 | 58.97 | 7.88 | 58.76 | 65.87 | 61.44 | 71.22 | 74.31 | 17.66 |
| TENT [58] | 80.11 | 92.78 | 7.27 | 54.75 | 63.51 | 7.83 | 58.52 | 58.05 | 7.47 | 67.36 | 74.11 | 48.66 | 72.60 | 76.67 | **15.88** |
| CoTTA [60] | 80.26 | 93.32 | 6.82 | 56.49 | 64.58 | 10.65 | **62.93** | 65.72 | 7.00 | **68.81** | 73.99 | 46.54 | 71.73 | 76.65 | 17.78 |
| SAR [43] | 81.15 | 92.67 | 7.80 | 55.04 | 63.77 | 8.28 | 58.55 | 58.16 | 7.38 | 66.86 | 73.24 | 50.6 | 72.75 | 76.82 | 16.11 |
| DIGA [61] | 83.76 | 92.69 | 8.93 | 55.82 | 64.33 | 9.34 | 58.77 | 60.72 | 6.91 | 67.35 | 73.74 | 52.04 | 69.27 | 75.19 | 17.89 |
| DomainAdaptor [72] | 78.01 | 90.82 | 7.37 | 37.31 | 46.80 | **6.84** | 58.57 | 58.53 | 7.28 | 65.49 | 72.36 | 54.78 | 72.38 | 76.36 | 17.00 |
| MedBN [44] | 80.35 | 92.16 | 11.75 | 51.54 | 59.16 | 7.84 | 58.58 | 58.55 | 8.19 | 57.41 | 63.26 | 60.06 | 70.76 | 73.86 | 18.00 |
| VPTTA [3] | 79.68 | 92.07 | 7.40 | 53.16 | 61.53 | 7.44 | 58.55 | 58.09 | 7.47 | 66.77 | 73.23 | 50.70 | 72.81 | 76.76 | 16.00 |
| **TopoTTA (Stage2 only)** | **88.63** | **96.62** | **4.77** | **57.45** | **66.70** | 8.27 | 62.88 | **71.28** | **5.72** | 68.50 | **75.55** | **37.32** | **74.03** | **77.51** | 17.33 |

Table C.3. The paired t-test results of clDice using UNet as baseline network across ten datasets. Source Only: Trained on the source, and tested on the target domain directly.

| Method | $p$ value | | | | |
|---|---|---|---|---|---|
| | DRIVE → CHASE | DRIVE → STARE | CHASE → DRIVE | CHASE → STARE | DeepGlobe → MR |
| Source Only | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| TENT [58] | 0.0011 | 0.0045 | <0.001 | <0.001 | <0.001 |
| CoTTA [60] | 0.0128 | <0.001 | <0.001 | 0.0033 | 0.001 |
| SAR [43] | 0.001 | 0.0085 | <0.001 | <0.001 | <0.001 |
| DIGA [61] | <0.001 | 0.057 | <0.001 | 0.027 | <0.001 |
| DomainAdaptor [72] | 0.0011 | <0.001 | <0.001 | <0.001 | <0.001 |
| MedBN [44] | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| VPTTA [3] | 0.0016 | 0.0032 | <0.001 | <0.001 | <0.001 |

| Method | $p$ value | | | | |
|---|---|---|---|---|---|
| | DeepGlobe → CNDS | Neub1 → Neub2 | Neub2 → Neub1 | ROSE → OCTA500 | OCTA500 → ROSE |
| Source Only | <0.001 | <0.001 | <0.001 | <0.001 | <0.001 |
| TENT [58] | <0.001 | 0.0036 | <0.001 | <0.001 | <0.001 |
| CoTTA [60] | <0.001 | 0.0146 | 0.0511 | <0.001 | 0.0067 |
| SAR [43] | <0.001 | 0.0053 | <0.001 | <0.001 | <0.001 |
| DIGA [61] | <0.001 | 0.0325 | 0.037 | <0.001 | <0.001 |
| DomainAdaptor [72] | <0.001 | 0.0015 | <0.001 | <0.001 | <0.001 |
| MedBN [44] | <0.001 | 0.0073 | <0.001 | <0.001 | 0.0013 |
| VPTTA [3] | <0.001 | 0.0128 | <0.001 | <0.001 | <0.001 |

## C.2. Additional Ablation Study

Due to the space limitation of our manuscript, we present comprehensive ablation study in this section.

**Impact of Different Iterations.** In the manuscript, we set the number of iterations to six. Here, we explore the impact of different iteration counts. Table C.4 shows the performance of TopoTTA and the competing methods under varying iteration settings. DIGA, which lacks a backforward process, is excluded from this comparison. TopoTTA consistently achieves the best performance across all iteration counts. For competing methods that update only BN parameters or external parameters, performance remains stable and shows strong robustness to iteration changes. Similar to TopoTTA, CoTTA demonstrates continuous performance improvement as the number of iterations increases. However, excessive iterations result in prolonged inference times. To balance performance and efficiency, we select six as the optimal number of iterations.

**Impact of Different Values of Hyperparameters.** To verify the sensitivity of our method to hyperparameters, we test the network performance under different hyperparameter conditions. As shown in Fig. C.3, we analyze the $s$ (modification window size in TopoHG), $n \times n$ (number of TopoMDCs regions), $k$ (coefficient for selecting the number of key points) and $\tau^{\text{bg}}$(upper limit of the foreground pixel ratio). From Fig. C.3(a), we can observe Dice initially rises and then drops sharply, while clDice stabilizes after reaching a peak. This is because small modification areas are insufficient to create *pseudo-*

Table C.4. Average testing results across four scenarios under different iteration counts. The best and second-best results in each column are highlighted in **bold** and underline, respectively.

| Method | Iterations = 2 | | | Iterations = 4 | | | Iterations = 6 | | | Iterations = 8 | | | Iterations = 10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice ↑ | clDice (%) ↑ | $\beta$ ↓ | Dice (%) ↑ | clDice (%) ↑ | $\beta$ ↓ |
| TENT [58] | 63.94 | 67.44 | 49.44 | 63.88 | 67.38 | 49.36 | 63.81 | 67.29 | 49.42 | 63.73 | 67.19 | 49.37 | 63.62 | 67.07 | 49.34 |
| CoTTA [60] | 64.99 | 68.92 | 46.82 | 65.33 | 69.11 | 47.31 | 65.49 | 69.34 | 47.30 | 66.62 | 70.45 | 47.06 | 67.55 | 71.46 | 46.95 |
| SAR [43] | 63.97 | 67.48 | 49.38 | 63.96 | 67.47 | 49.46 | 63.97 | 67.49 | 49.32 | 63.95 | 67.45 | 49.52 | 63.93 | 67.42 | 49.55 |
| DomainAdaptor [72] | 63.41 | 66.53 | 49.98 | 63.42 | 66.53 | 49.98 | 63.41 | 66.53 | 49.98 | 63.41 | 66.53 | 49.98 | 63.41 | 66.53 | 49.98 |
| MedBN [44] | 56.21 | 53.53 | 83.75 | 56.57 | 54.42 | 82.22 | 56.65 | 54.90 | 80.05 | 56.61 | 54.88 | 79.78 | 56.73 | 54.97 | 76.75 |
| VPTTA [3] | 63.98 | 67.48 | 49.75 | 63.98 | 67.48 | 49.79 | 63.98 | 67.48 | 49.76 | 63.99 | 67.49 | 49.76 | 63.99 | 67.49 | 49.76 |
| **TopoTTA (Ours)** | **67.72** | **71.51** | **43.90** | **68.48** | **73.22** | **42.38** | **69.44** | **74.00** | **43.01** | **69.31** | **74.33** | **42.93** | **69.32** | **74.62** | **42.29** |

*breaks*, whereas excessively large areas can result in falsely high continuity due to overly aggressive predictions. As shown in Fig. C.3(b), both Dice and clDice achieve their highest values when $4 \times 4$ is used. When $n$ is too small, TopoMDCs struggle to capture the varied topological features across regions, leading to lower performance. Conversely, when $n$ is too large, the increased number of learnable $\boldsymbol{\delta}$ parameters makes learning more difficult, also reducing performance. From Fig.C.3(c), we select 0.002 as the optimal value for $k$ to achieve the best overall performance. Similarly, Fig. C.3(d) shows that Dice and clDice also reach their peak at the same value. Including too many foreground pixels in the background window hinders the effective creation of *pseudo breaks*, while overly strict constraints on the foreground pixel ratio result in an insufficient number of *pseudo breaks*.

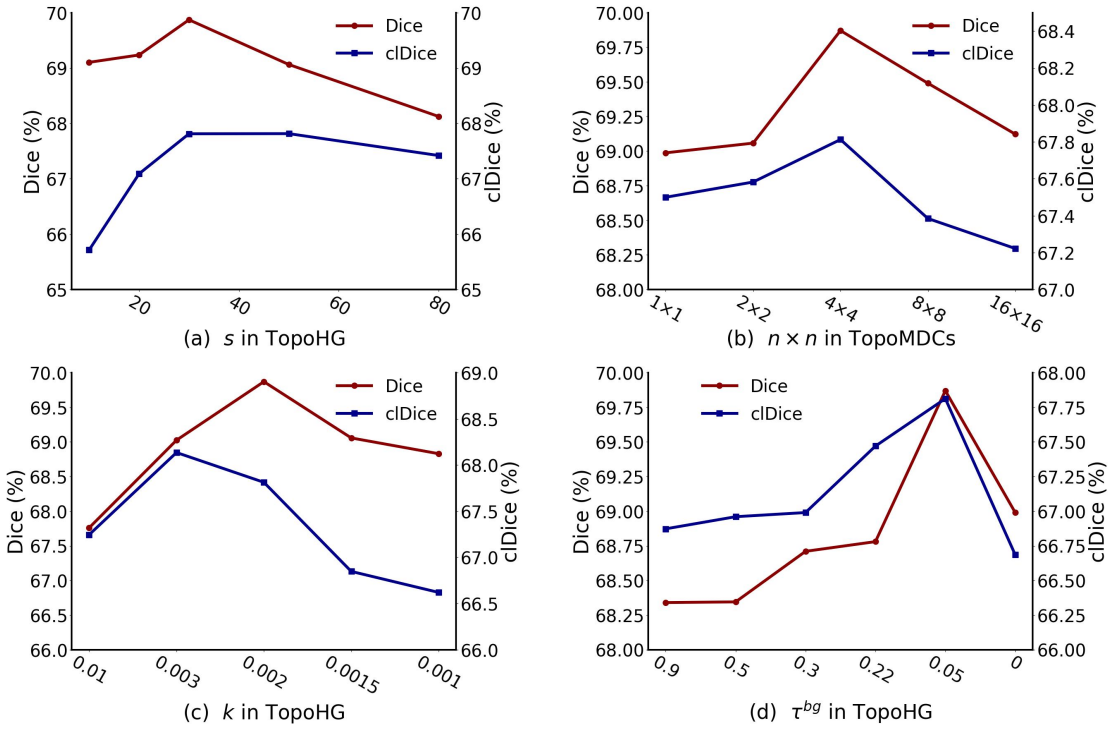


Figure C.3. Performance of TopoTTA with different hyperparameter conditions.

**Impact of different synthesis qualities.** Our background patch search strategy around neighbors, combined with low-frequency swap, ensures sample authenticity at most times. To further validate it, we conduct experiments with background patches of varying similarity, where higher similarity corresponds to better synthesis quality. Specifically, we compute the similarity between all patches in the image and the selected foreground patch. We then conduct low-frequency swap using the least similar, moderately similar, and most similar ones, respectively, and evaluate their individual performances. As shown in Table C.5, the performance of using neighbor patch (Ours) is almost identical to that of using the best quality patches (need extra time +3.02s per image), indicating the high quality of our synthesized samples. To trade off time and accuracy, our method remains the preferred choice. Note that even with the worst quality patches, performance still outperforms the second-best baseline.

Table C.5. Ablation results of the different synthesis qualities. The best and second-best results in each column are highlighted in **bold** and underline, respectively.

| Synthesis quality | DRIVE → CHASE | | | CHASE → DRIVE | | |
|---|---|---|---|---|---|---|
| | Dice (%) ↑ | clDice (%) ↑ | $\beta \downarrow$ | Dice (%) ↑ | clDice (%) ↑ | $\beta \downarrow$ |
| CoTTA | 68.60 | 71.53 | 36.38 | 67.64 | 64.80 | 81.20 |
| Worst quality (least similar) | 69.02 | 73.82 | 26.13 | 71.75 | 67.65 | 84.05 |
| Middle quality (moderately similar) | 70.69 | 75.88 | 25.87 | 72.55 | 68.85 | 82.00 |
| Best quality (most similar) | 70.23 | **77.65** | **23.35** | **73.22** | **70.46** | 80.20 |
| **Neighbor** | **70.73** | 77.05 | 25.38 | 72.96 | 70.26 | **79.15** |

**Necessity of updating router parameter.** To verify that simultaneously adjusting both $\delta$ and the model parameters increases the search-space complexity, we conduct an experiment where all parameters are updated together. As shown in Table C.6, the results reveal a significant performance drop, indicating that this approach may introduce greater parameter instability. And in the paper's setting, router parameters $\delta$ adds only 1280 params, a negligible increase compared to the original model's params ($2.894 \times 10^6$), and fewer than methods like VPTTA, which add 4332 params.

Table C.6. Ablation results of adjusting either all parameters or only router parameters $\delta$ at Stage 1.

| Upadate params | Param size | DRIVE → CHASE | | | CHASE → DRIVE | | |
|---|---|---|---|---|---|---|---|
| | | Dice (%) ↑ | clDice (%) ↑ | $\beta \downarrow$ | Dice (%) ↑ | clDice (%) ↑ | $\beta \downarrow$ |
| All | $2.895 \times 10^6$ | 70.07 | 72.58 | 30.23 | 69.08 | 65.77 | 81.7 |
| $\delta$ | 1280 | **70.73** | **77.05** | **25.38** | **72.96** | **70.26** | **79.15** |

**Visualization of Different Data Augmentation Methods.** We visualize the effects of different data augmentation methods, as shown in Fig. C.4. Our frequency-based method effectively preserves high-frequency details in the generated *pseudo-breaks*. In comparison, the Gaussian blur method shows moderate visual effects, while random Gaussian noise and image swap methods overly modify the images, creating actual breaks.

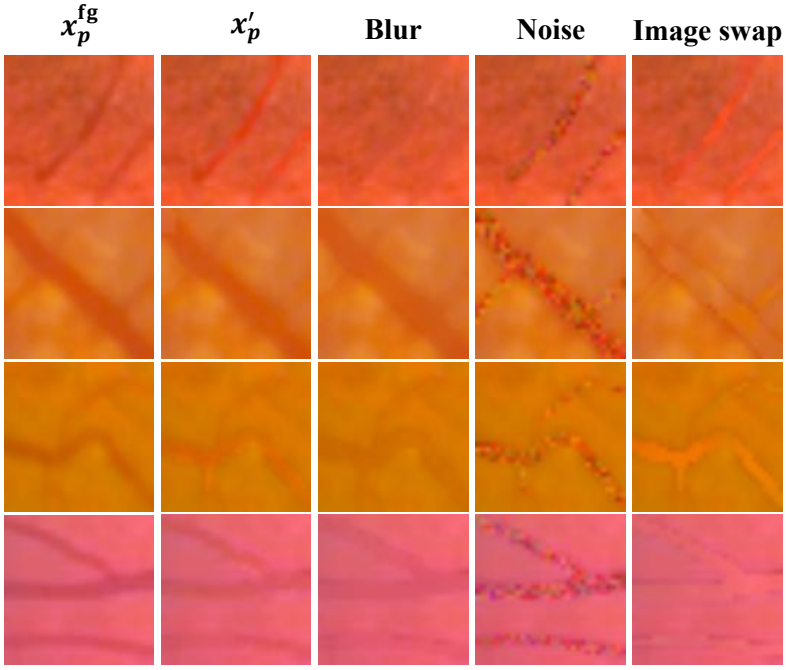


Figure C.4. Visualizations of pseudo-breaks generated by TopoHG and three data augmentation methods, i.e., Gaussian blur, random Gaussian noise, and image swap in the spatial domain. $\boldsymbol{x}_p^{\text{fg}}$ is original patch and $\boldsymbol{x}'_p$ denotes pseudo-break generated by TopoHG.

## C.3. Additional Visualization Results

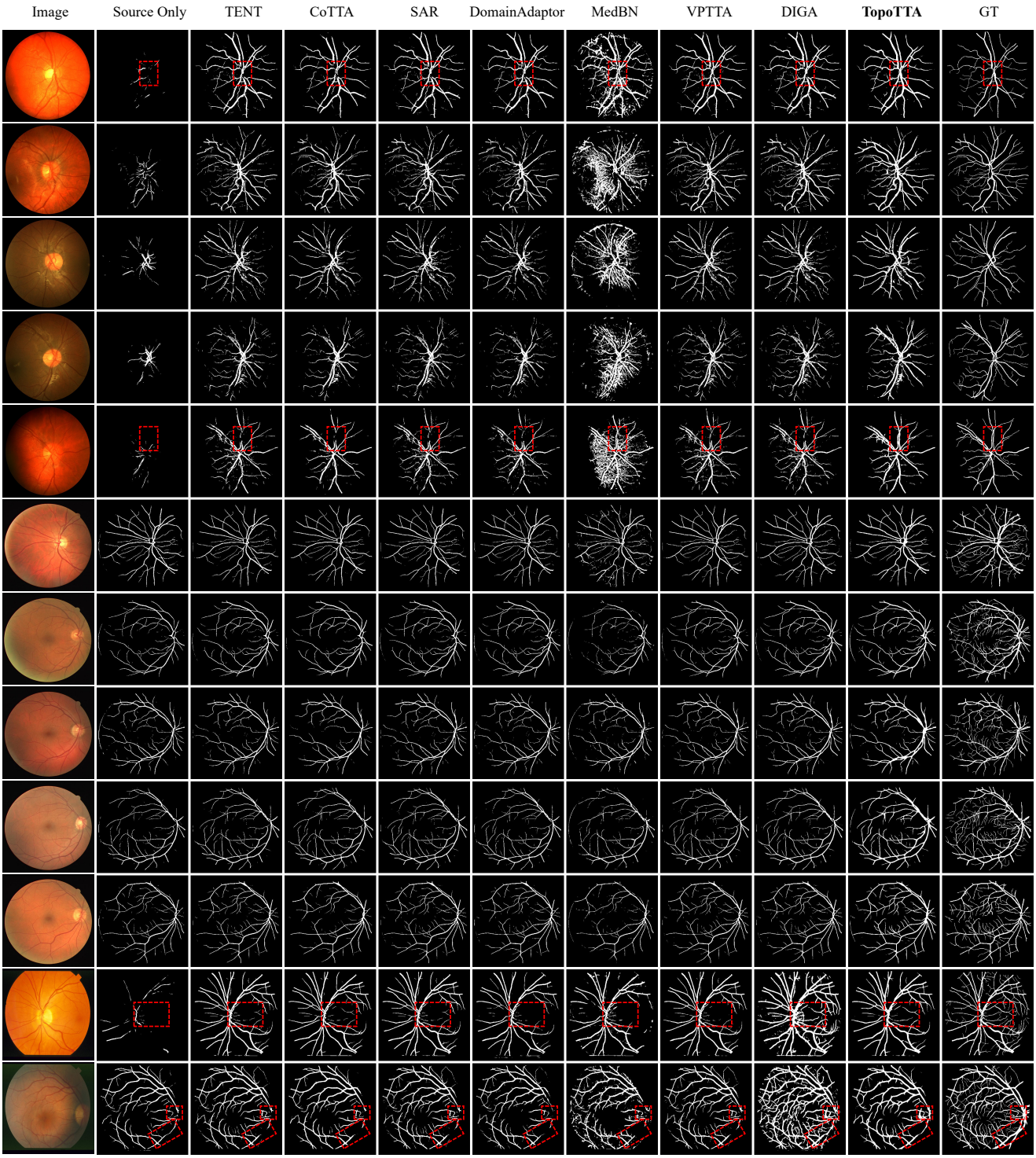


Figure C.5. Visualization of segmentation results for TopoTTA and seven comparison methods in retinal vessel segmentation scenario. "Source Only" denotes results without any TTA methods applied, and GT is short for ground-truth labels.

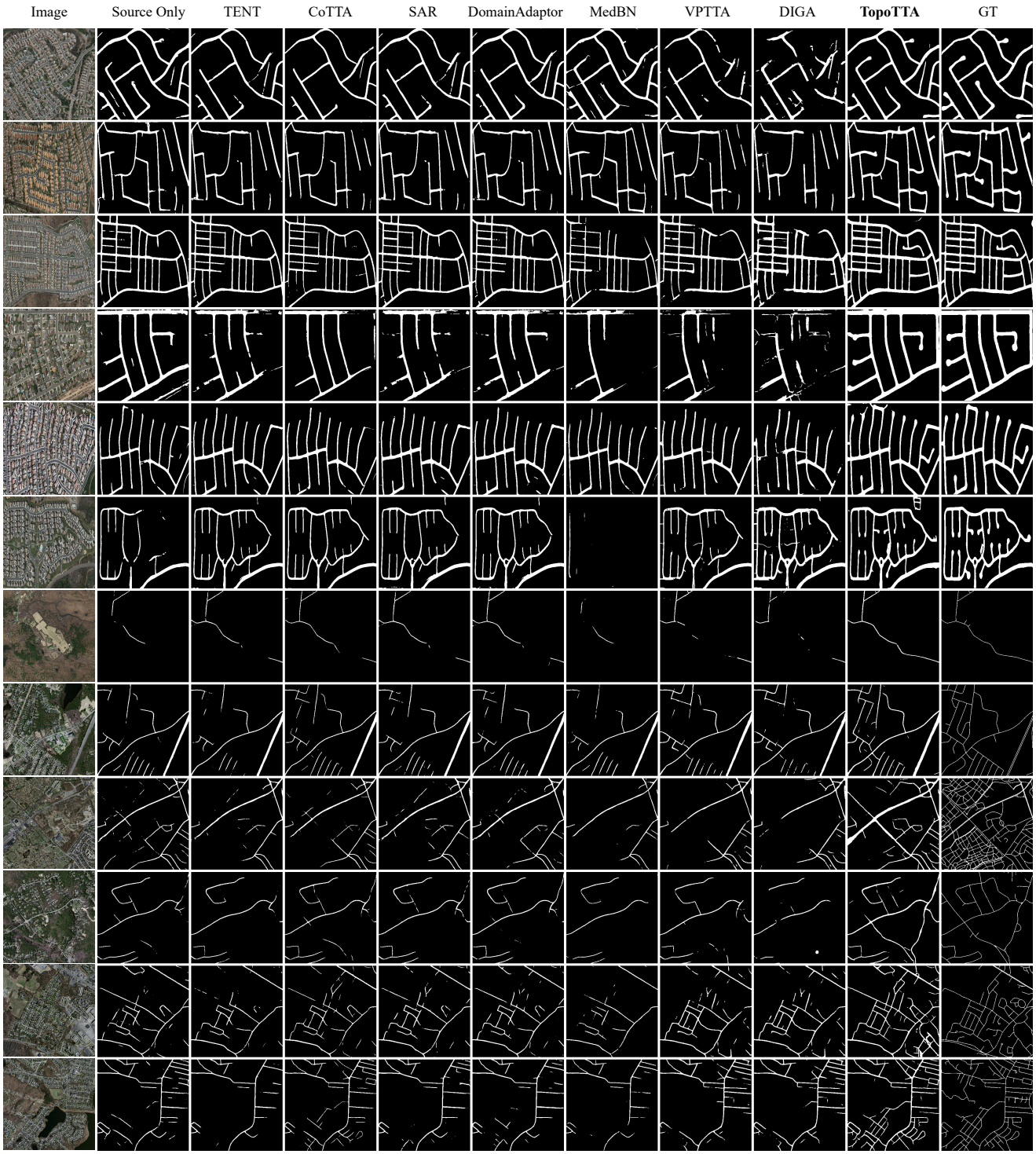


Figure C.6. Visualization of segmentation results for TopoTTA and seven comparison methods in road extraction scenario. “Source Only” denotes results without any TTA methods applied, and GT is short for ground-truth labels.

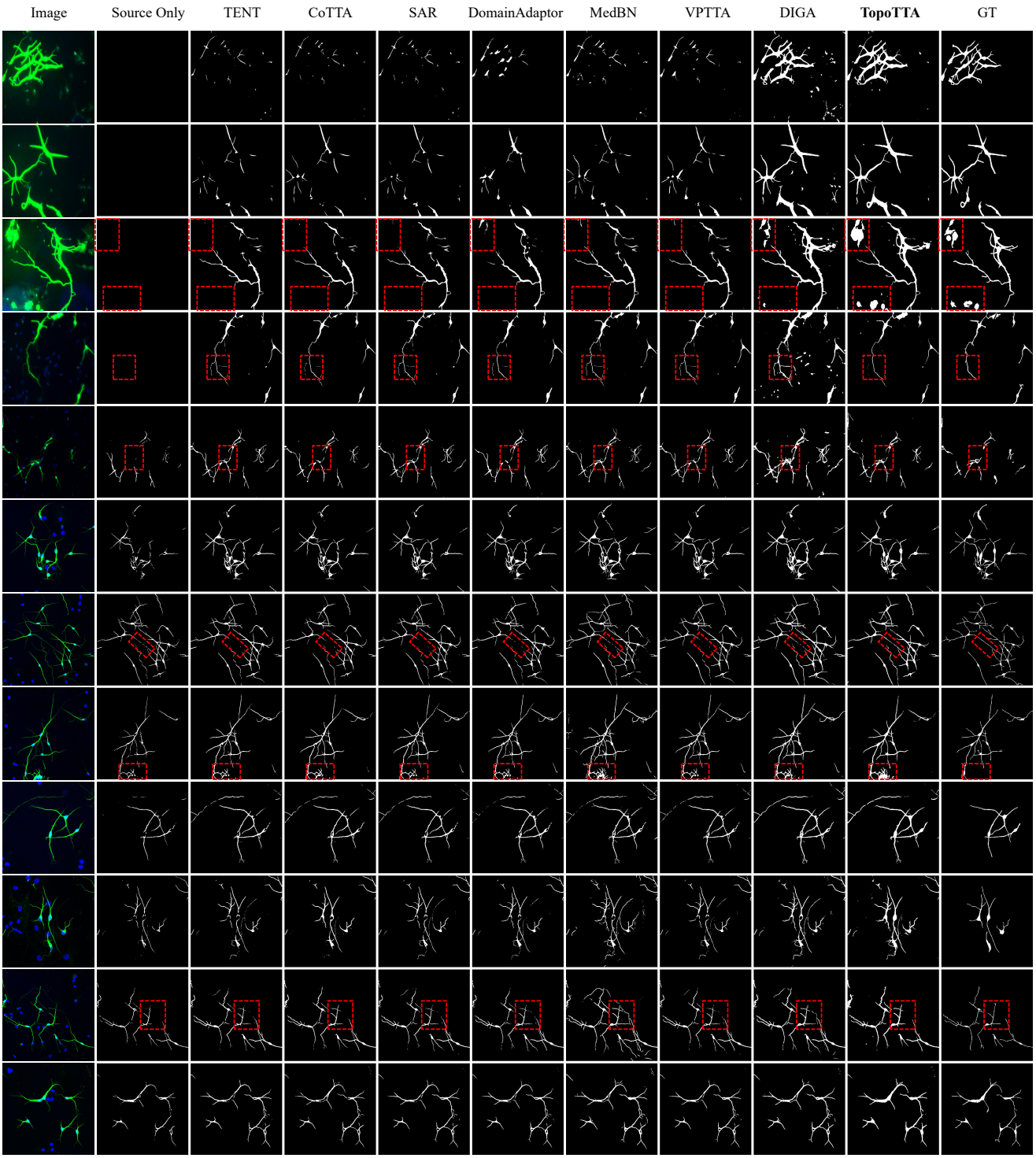


Figure C.7. Visualization of segmentation results for TopoTTA and seven comparison methods in microscopic neuronal segmentation scenario. “Source Only” denotes results without any TTA methods applied, and GT is short for ground-truth labels.

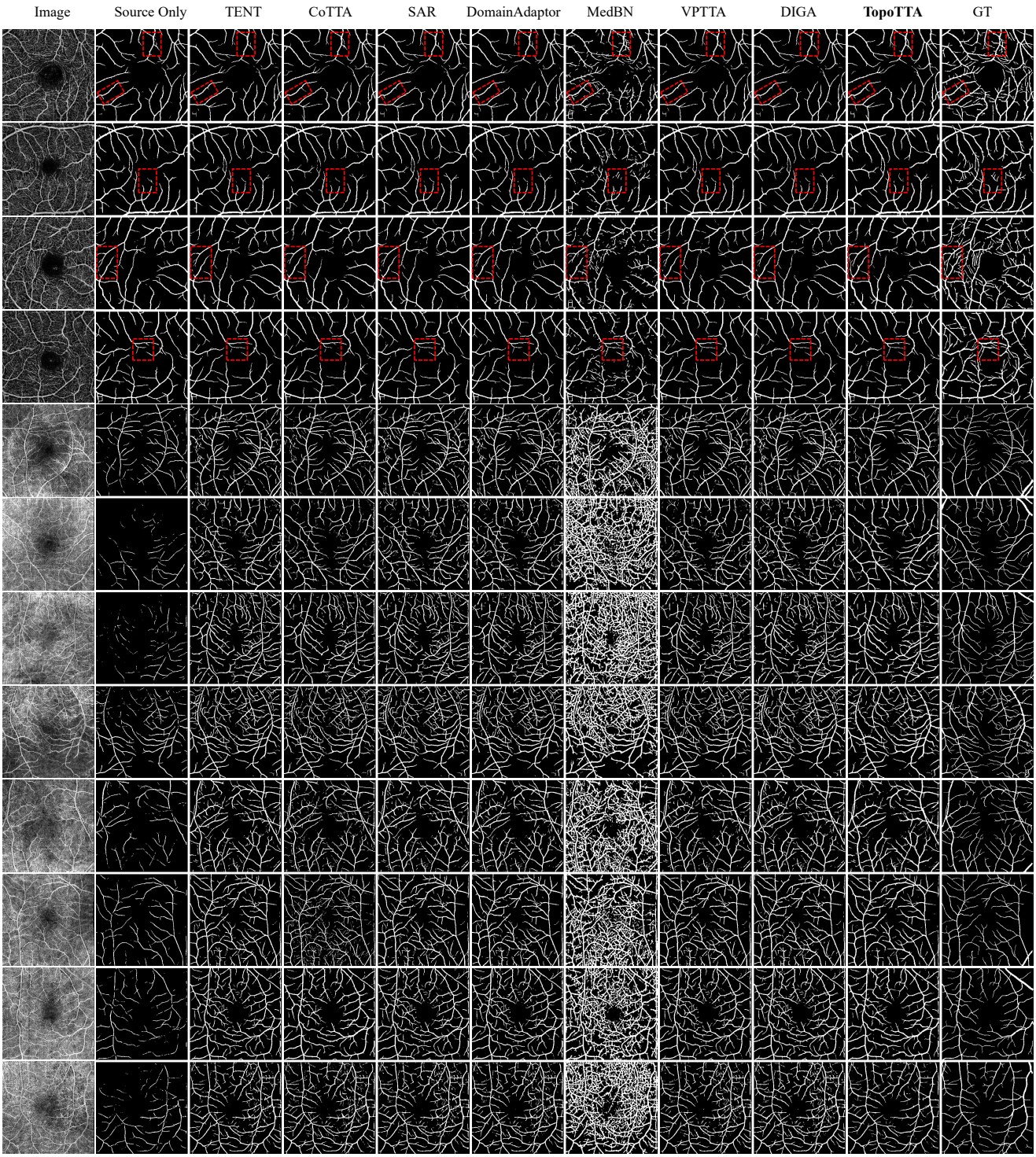


Figure C.8. Visualization of segmentation results for TopoTTA and seven comparison methods in retinal OCT-angiography vessel segmentation scenario. “Source Only” denotes results without any TTA methods applied, and GT is short for ground-truth labels.