Supplementary materials for
BlindHarmony: "Blind" Harmonization for MR Images via Flow model

1. Dataset description

The BlindHarmony framework proposed in this work was trained and evaluated using the OASIS3 [2], employing a target domain consisting of images acquired with a Siemens TIM Trio 3T MR scanner. Meanwhile, as a source domain, images obtained from four other scanners were utilized, including the Siemens Magnetom Vida 3T MR scanner (Domain 1), the Siemens Vision 1.5T scanner (Domain 2), the Siemens BioGraph mMR PET-MR 3T scanner (Domain 3), and the Siemens TIM Trio 3T MR scanner (Domain 4). Domain 4 scanner shared the same scanner version with the target domain scanner. All images were resampled to a uniform resolution of $1.2\times1.2\times1.2$ mm and normalized using min-max normalization at the slice level. Table 1 provides detailed information regarding the acquisition parameters.

2. Training details

2.1. Flow model

For our experiments, we employed the Neural Spline Flow (NSF) architecture [1] with rational quadratic (RQ) spline coupling layers. The majority of the hyperparameters for the network were set to the same values used in the original NSF paper for experiments on the ImageNet dataset. Specifically, the tail bound B and the number of bins K for the RQ spline coupling layers were set to B = 3 and K = 8, respectively. A multiscale architecture similar to that of Glow was utilized, with each layer of the network consisting of 7 transformation steps, including an actnorm layer, an invertible 1×1 convolution, an RQ spline coupling transform, and another 1×1 convolution. The network itself comprises 4 layers, which results in a total of 28 coupling transformation steps. Additionally, 3 residual blocks and batch normalization layers are included in the subnetworks parameterizing the RQ splines. An Adam optimizer with an initial learning rate of 0.0005 and cosine annealing of the learning rate was used to iteratively optimize the parameters up to 20K steps. Two separate models were trained, one for simulated dataset validation and one for real-world dataset validation. The same hyperparameters were assigned to both models except for the dataset composition, with the simulated evaluation dataset consisting of 76775/775/800 slices for training/validation/testing, and the real-world evaluation dataset consisting of 75240/760/1000 slices for training/validation/testing. An example of a sampled image of the flow model is illustrated in Figure 1.

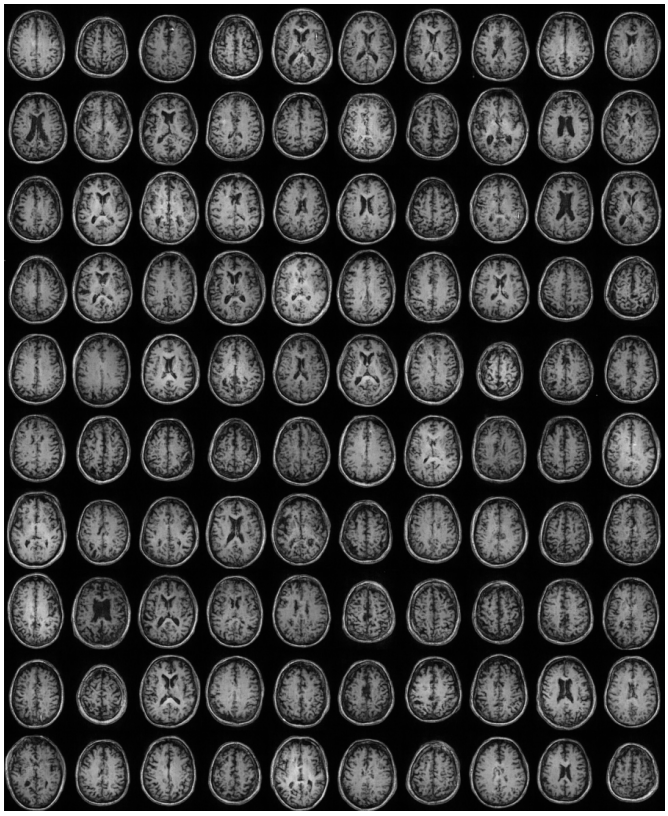

Figure 1. Example images of the trained flow model.

2.2. U-Net

The U-Net [3] architecture includes 4 Down blocks and 4 Up blocks, each block consisting of two sequences of convolution layer, batch normalization, and ReLU activation. The Down blocks utilize maxpooling, while the Up blocks use bilinear upsampling. The U-Net also utilizes skip connections. The training was done with 120 epochs using L1

	Target domain	Domain 1	Domain 2	Domain 3	Domain 4
Manufacturer	Siemens	Siemens	Siemens	Siemens	Siemens
scanner version	TIM Trio	Magnetom Vida	Vision	BioGraph mMR	TIM Trio
Magnetic field strength (T)	3	3	1.5	3	3
Matrix size	176×256×256	176×240×256	128×256×256	176×240×256	176×256×256
resolution (mm)	1×1×1	1.2×1.05×1.05	1.25×1×1	1.2×1.05×1.05	1×1×1
TR/TI (s)	2.4/1	2.3/unknown	9.7/unknown	2.3/0.9	2.4/1
TE (ms)	3.2	3.0	4	3.0	3.2
Flip angle (°)	8	9	10	9	8
Total number of sessions	1568	378	620	879	625

Table 1. Scan parameters of domains in OASIS3 dataset is illustrated.

	Domain 1	Domain 2	Domain 3	Domain 4
U-net (slices)	5100/1000	9750/600	25150/1000	19350/1000
CylceGAN (slices)	17400/1000	29000/600	42100/1000	29050/1000

Table 2. The number of training/test datasets for each domain.

loss and the Adam optimizer with a learning rate of 0.001. For the simulated dataset evaluation, the same dataset composition used in training the flow model was utilized, with a composition of train/val/test = 76775/775/800 slices. For the real-world dataset application, the validation step was dropped to increase the number of training data. The number of datasets varies by domain, as described in Table 2.

2.3. CycleGAN

The generator in CycleGAN [4] consists of 2 convolution layers with instance normalization and ReLU activation, 9 residual blocks, and 3 convolution layers with instance normalization and ReLU activation. Each residual block includes a residual connection of 2 convolution layers with instance normalization and ReLU activation. The discriminator consists of 5 convolutional layers, 4 leaky ReLU, and 4 instance normalization. The CycleGAN uses identity loss, cycle loss, and adversarial loss, and training was done with 40 epochs using the Adam optimizer with a learning rate of 0.0002. For the simulated dataset evaluation, the same dataset composition used in training the flow model was utilized, with a composition of train/test = 76775/775/800 slices. For the real-world dataset application, the validation step was dropped as in the U-Net cases, and the number of target domain images was train/test = 75240/1000 slices, while the number of source domain images varied by domain, as described in Table 2.

3. Ablation study

We compared the PSNR and SSIM values of a simulated dataset to a baseline by varying one of the hyperparameters α, β_1, and β_2.

PSNR	Exp	Log	Gamma 0.7
Baseline	29.6	28.8	27.4
$\alpha = 0$	29.5	28.5	27.2
$\beta_1 = 0$	18.5	18.5	18.6
$\beta_2 = 0$	29.6	28.8	27.4

SSIM	Exp	Log	Gamma 0.7
Baseline	0.985	0.978	0.969
$\alpha = 0$	0.984	0.977	0.968
$\beta_1 = 0$	0.693	0.695	0.696
$\beta_2 = 0$	0.985	0.978	0.969

Table 3. The ablation study.

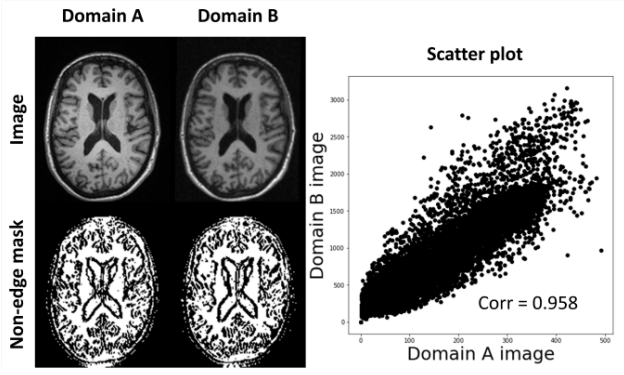

Figure 2. Visual illustration for Eq. 1 and Eq. 2.

4. Illustration for Eq. 1 and Eq. 2

When harmonizing two T_1 weighted images from different scanners, the images are primarily created by the prop-

erties of the brain (*e.g.*, T1, proton density) while scanner-specific differences are mostly spatially slow-varying patterns (*e.g.*, center-brightening from B1 inhomogeneity). Consequently, a high spatial correlation (Eq. 1) exists between the images, and their edges are likely to coincide (Eq. 2). These can be observed in Fig. 2. The non-edge masks of the images from the two different domains exhibit a high coincidence, and the scatter plot demonstrates a strong correlation between the two images. If these conditions are not met, it may create a failure case (e.g., harmonizing T_1 and T_2 weighted images).

5. Visual examples

Exemplary images of the simulated source domain dataset application, real-world data application, and the downstream task application are illustrated in Figures 3, 4 and 5

References

[1] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv*, pages 2019–12, 2019.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention, Part III 18*, pages 234–241. Springer, 2015.

[4] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2223–2232, 2017.

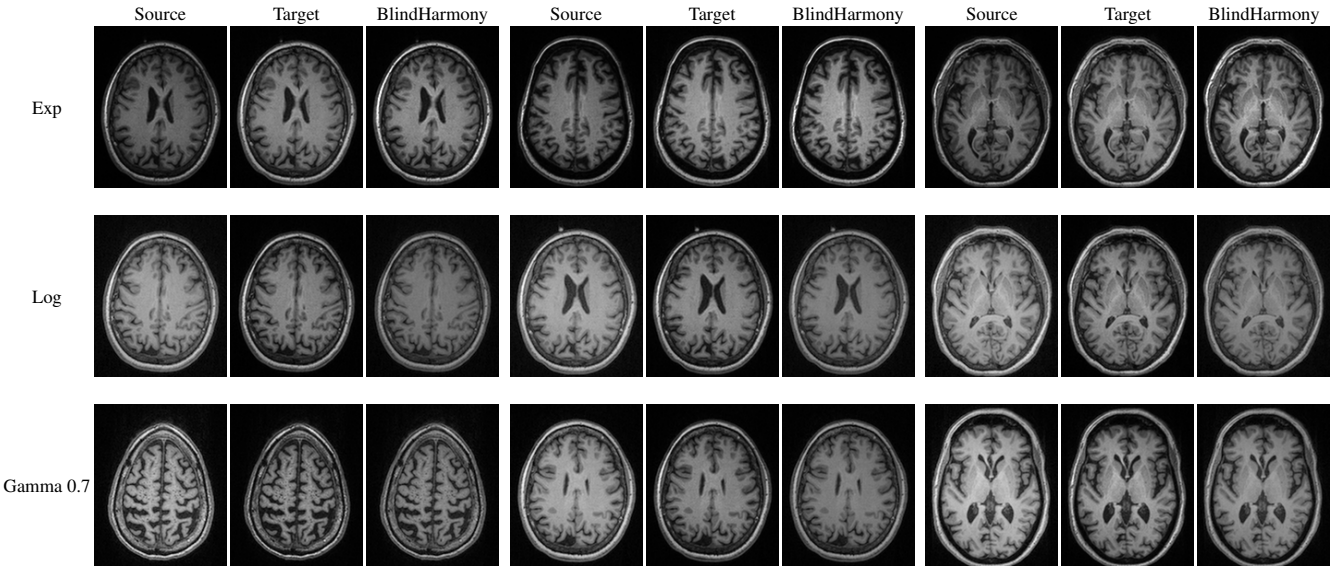

Figure 3. Example images of BlindHarmony application of simulated source domain images.

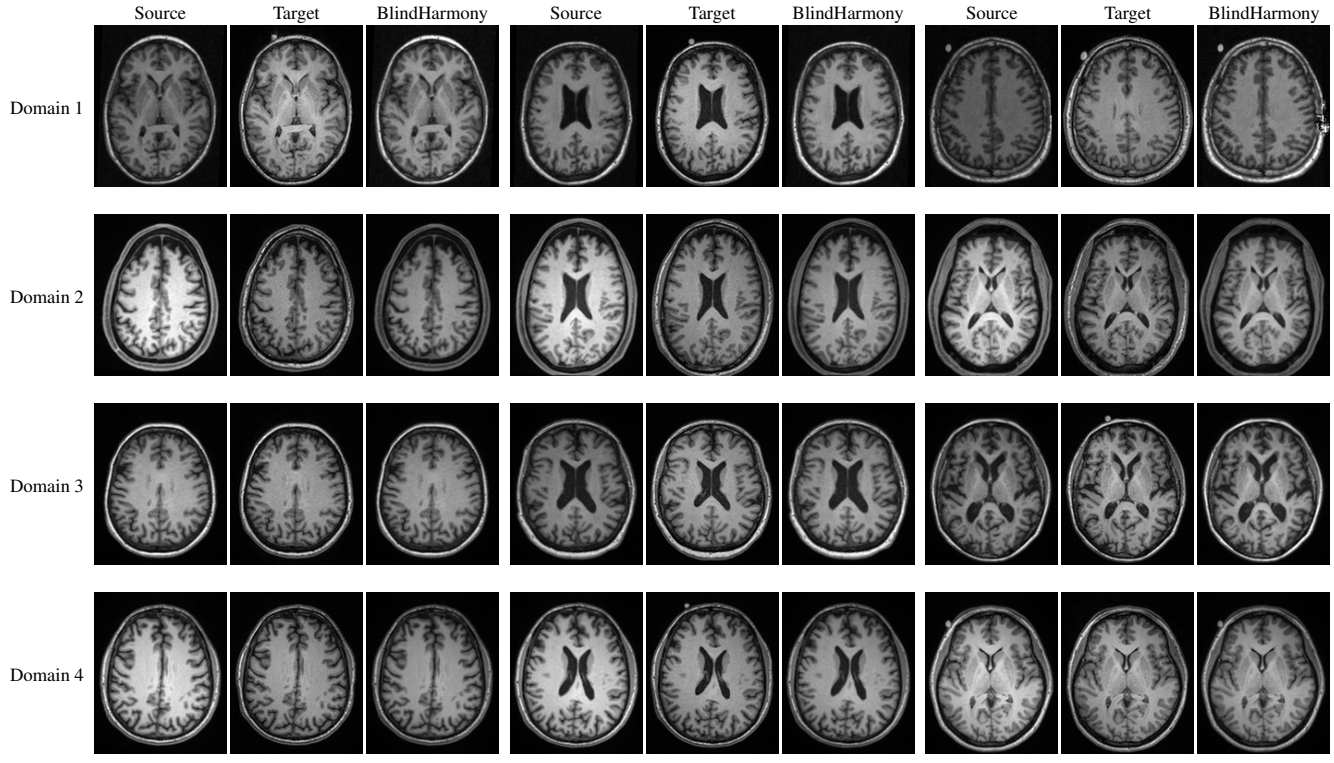

Figure 4. Example images of BlindHarmony application of real-world data images.

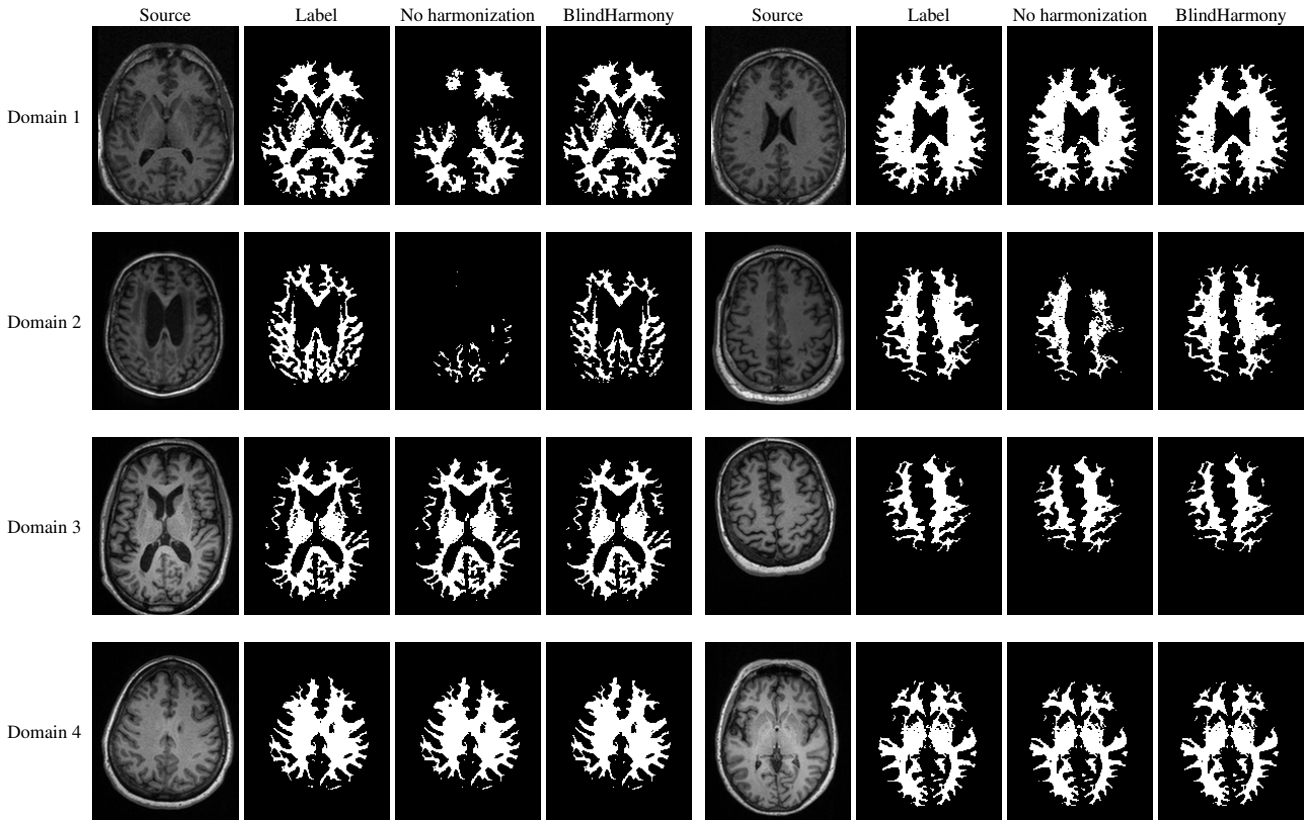

Figure 5. Example images of BlindHarmony application to the downstream task of white matter segmentation network.