Modality-Adaptive Contrastive and Supervised Learning for Multi-Field MRI-to-7T Synthesis

Junsheng Mi[1]

Xi'an Jiaotong University, Xi'an, Shaanxi, China

Abstract. Synthesizing ultra-high-field magnetic resonance images from routinely acquired or low-field scans is challenging because the appearance gap depends jointly on pulse sequence and source field strength, while paired 7T data are scarce. We present our submission to Task 1 of the MRIxFields 2026 challenge, which treats T1-weighted (T1W), T2-weighted (T2W), and T2-FLAIR images with complementary learning strategies. T1W translation uses field-specific contrastive unpaired translation followed by structure-aware paired fine-tuning and late-checkpoint averaging. T2W uses one field-conditioned supervised residual generator. For T2-FLAIR, a supervised prediction is blended with a field-specific contrastive prediction. A low-capacity teacher-anchored stage further adapts only the last residual blocks and decoder while constraining predictions on unpaired retrospective scans. On the challenge validation server the submitted system obtains mean SSIM 0.9095, nRMSE 0.2955, LPIPS 0.0776, and matched Dice 0.8809. A controlled comparison supports the modality-adaptive design: one uniform field-conditioned supervised generator applied to T1W reaches SSIM 0.9594 against 0.9812 for our contrastive branch, losing at every source field. Leave-one-out experiments show teacher anchoring improves all twelve held-out subject–field combinations, though only by order 10^{-4}, whereas augmentation and a stand-alone 2.5D model do not generalize. We further quantify that background masking does not inflate our metrics, and report containerized inference at 15.6 s per case.

Keywords: MRI synthesis · ultra-high-field MRI · contrastive learning · conditional generation · domain adaptation

1 Introduction

Ultra-high-field 7T MRI offers high signal-to-noise ratio and conspicuous anatomical detail, but availability, acquisition time, contraindications, and protocol heterogeneity limit its routine use, making synthesis from lower fields attractive for harmonization and downstream analysis. This is not a single image-to-image mapping: a 0.1T scan differs from 7T far more strongly than a 5T scan, and T1W, T2W, and T2-FLAIR exhibit different contrast relationships. Paired conditional adversarial networks work when source and target are aligned [2, 5], whereas cycle-consistent and contrastive translation exploit larger unpaired collections [3, 4]; shared latent representations are another route [6]. In MRIxFields the prospective paired set contains only three subjects per source field while retrospective data provide many more unpaired images, so high-capacity paired adaptation risks memorizing those three subjects.

This is a challenge-system report: its contribution is not a new learning principle but an empirical demonstration of which supervision suits which modality under scarce pairing, with the measurements that bound that claim. We assign contrastive field-specific translation to T1W, conditional supervised translation to T2W, and prediction-level fusion to T2-FLAIR; show that one uniform supervised generator is markedly worse on T1W at every source field; add teacher-anchored fine-tuning reported with per-subject confidence intervals; and give a reproducible inference protocol with measured cost, including an ablation showing our masking does not inflate the metrics. Negative 2.5D, augmentation, and calibration results are also reported.

2 Method

2.1 Problem Formulation and Shared Preprocessing

Let $x_{m,f}$ denote a source image of modality $m \in \{\mathrm{T1W}, \mathrm{T2W}, \mathrm{T2FLAIR}\}$ and field strength $f \in \{0.1, 1.5, 3, 5\}$T, with $y_{m,7}$ its 7T target when available. Each volume is converted to canonical RAS+ orientation and normalized to $[0, 1]$; axial slices at native 364×436 resolution are mapped to $[-1, 1]$ before entering a generator, with no crop or resize. Predictions are mapped back to $[0, 1]$, restored to the input orientation, and multiplied by $\mathbb{1}[x > 10^{-6}]$. This masking is a determinism safeguard, not a source of accuracy: Section 4.4 shows it changes mean SSIM by -1.8×10^{-4}. We retain it because it guarantees exact zeros outside the source support, making output validation deterministic.

All generators use a nine-block residual architecture [10] with 64 base channels: 11,365,633 parameters unconditionally, 11,378,177 with field conditioning. The T2W and T2-FLAIR models are independently trained instances of that conditional architecture and share no parameters. Figure 1 summarizes the branches.

2.2 T1W: Contrastive Translation and Structural Fine-Tuning

For each source field we first train a separate Contrastive Unpaired Translation (CUT) model [4] on retrospective source and 7T domains. PatchNCE features are sampled from generator layers $\{0, 4, 8, 12, 16\}$ with temperature 0.07 and 256 patches; adversarial and NCE terms have unit weight. Pretraining runs 100 epochs with batch size 2 and learning rate 2×10^{-4}. The generator is then paired-fine-tuned for 100 epochs using

$$\mathcal{L}_{\mathrm{T1}} = \mathcal{L}_1 + 0.1\mathcal{L}_{\mathrm{LPIPS}} + \mathcal{L}_{\mathrm{SSIM}} + 0.5\mathcal{L}_{\mathrm{edge}}, \tag{1}$$

where $\mathcal{L}_{\mathrm{SSIM}} = 1 - \mathrm{SSIM}$ [7], LPIPS is a deep feature distance [8], and the edge term is the L_1 distance between horizontal and vertical Sobel responses, at

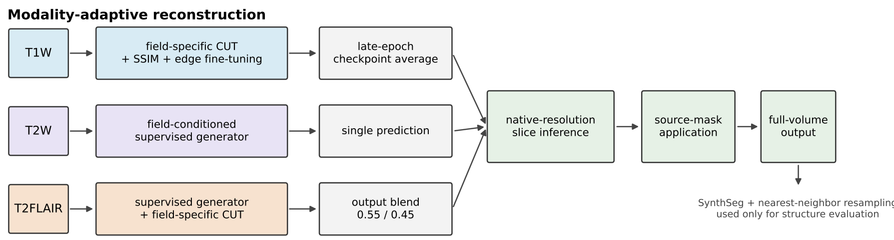

Fig. 1. Overview of the modality-adaptive reconstruction system. T1W uses field-specific contrastive models and checkpoint averaging; T2W uses one field-conditioned supervised model; T2-FLAIR fuses supervised and contrastive outputs. All branches share the same deterministic full-volume inference protocol.

learning rate 5×10^{-5}. Checkpoints are saved on a fixed 10-epoch grid, and the rule fixed in advance was to average the last saved ones rather than search over subsets: epochs 90/100 for 1.5T, 80/90/100 for 3T, 90/100 for 5T, contributing +0.0004 mean SSIM. At 0.1T the shipped model is instead the teacher-anchored student of Section 2.5. How many checkpoints to average per field was informed by server feedback (Section 4.4).

2.3 T2W: Field-Conditioned Supervised Synthesis

The paired T2W relationship is stable enough for a shared conditional model. A four-dimensional one-hot field code is broadcast spatially and concatenated with the input slice, giving five input channels. Training runs 120 epochs at batch size 4 and learning rate 2×10^{-4}, with adversarial term following the standard GAN formulation [1]:

$$\mathcal{L}_{\text{sup}} = 10\mathcal{L}_1 + 0.2\mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{LPIPS}} + 5\mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{edge}}. \tag{2}$$

The low adversarial weight discourages hallucinated texture, while the field code lets one network represent four source-to-7T mappings.

2.4 T2-FLAIR: Prediction-Level Hybridization

For T2-FLAIR the supervised model yields stable anatomy but underestimates some high-frequency 7T appearance, so we also fine-tune a CUT model per source field and blend predictions rather than weights:

$$\hat{y}_{\text{FLAIR}} = (1-\alpha)G_{\text{sup}}(x,f) + \alpha G_{\text{CUT},f}(x), \qquad \alpha = 0.45. \tag{3}$$

Prediction-level fusion lets the branches retain independent normalizations. The coefficient was chosen on the validation server, but the sweep is flat where it matters: $\alpha = 0.40$ and $\alpha = 0.45$ differ by 5×10^{-6} SSIM. The supportable conclusion is that fusion helps and any $\alpha \in [0.40, 0.45]$ is equivalent, not that 0.45 is optimal.

2.5 Teacher-Anchored Low-Capacity Adaptation

To reduce prospective overfitting we adapt the 0.1T T1W model with a frozen copy of the submitted generator as teacher, training only the last three residual blocks and decoder (3.91M of 11.37M parameters). Paired slices use $\mathcal{L}_1 + 0.3\mathcal{L}_{\mathrm{SSIM}} + 0.1\mathcal{L}_{\mathrm{edge}} + 0.1\mathcal{L}_{\mathrm{nRMSE}}$ with $\mathcal{L}_{\mathrm{nRMSE}} = \|\hat{y} - y\|_2 / \|y\|_2$. For each paired update a retrospective slice is perturbed by gain, gamma, and Gaussian noise, and the student must match the teacher's clean prediction via $L_1 + 0.2\mathcal{L}_{\mathrm{SSIM}}$. We add an L2-SP penalty [11] toward the initial student weights:

$$\mathcal{L}_{\mathrm{anchor}} = \mathcal{L}_{\mathrm{paired}} + 2\mathcal{L}_{\mathrm{teacher}} + 10^{-4}\|\theta - \theta_0\|_2^2. \tag{4}$$

This functional consistency is related to teacher-based regularization [12], but our teacher stays fixed. We train six epochs at learning rate 2×10^{-6} with batch size 2.

All three prospective subjects are used for paired fine-tuning and for the shipped anchored model. The leave-one-out (LOO) analysis of Section 4.3 instead trains on two subjects and evaluates on the third; those LOO models serve only that analysis, never the container.

2.6 Submitted System and Full-Volume Evaluation

Table 1 lists the submitted container: 14 checkpoints totalling 1.6 GB. Most benefit lies in the modality split rather than the ensembles: dropping the ensembles and anchoring would ship 6 checkpoints and forfeit roughly 0.0017 SSIM.

For development-time Dice and volume evaluation, SynthSeg [9] is applied to the complete reconstructed volume, and the label map is nearest-neighbor resampled to the 0.5-mm grid before extracting the scored slices $z \in [150, 180)$; we never segment the isolated slab.

Table 1. Complete inventory of the submitted Task 1 container.

Branch	Checkpoints	#	Role
T1W 0.1T	`0.1T_epoch6`	1	teacher-anchored student
T1W 1.5T	`1.5T_epoch{90,100}`	2	late-checkpoint ensemble
T1W 3T	`3T_epoch{80,90,100}`	3	late-checkpoint ensemble
T1W 5T	`5T_epoch{90,100}`	2	late-checkpoint ensemble
T2W	`T2W_netG_epoch120`	1	one model, all four fields
T2-FLAIR	`T2FLAIR_netG_epoch120`	1	supervised branch
T2-FLAIR	`{0.1,1.5,3,5}T_cut_epoch30`	4	contrastive branch
Total		**14**	1.6 GB on disk

3 Experiments

3.1 Data and Protocol

The challenge provides T1W, T2W, and T2-FLAIR volumes at 0.1T, 1.5T, 3T, 5T, and 7T. Each prospective training domain contains the same three paired subjects; retrospective domains contain 43–235 unpaired volumes. The hidden-label validation set has three subjects per source field and modality, for 36 reconstructed volumes of shape $364 \times 436 \times 364$ at 0.5 mm, of which the server scores 30 central slices. The primary metric is mean SSIM across the 12 subtasks; we also report nRMSE, LPIPS, Dice, and relative volume agreement. Anchor-experiment selection uses three-fold subject-level LOO over IDs 0006, 0007, and 0009. All experiments use PyTorch and RTX 4090 GPUs, with no test-time augmentation.

3.2 Baselines and Ablations

Our principal internal reference is the *matched single-checkpoint submission*: T1W using the field-specific CUT model at epoch 100 for all fields without averaging or anchoring, T2W and T2-FLAIR each using their supervised generator at epoch 120 without fusion, and the shared inference protocol. It is matched because reconstructions and segmentations were regenerated together, and is our own configuration, not an external baseline.

To test the central design claim we compare our T1W branch against one uniform field-conditioned supervised generator, matched in architecture and recipe: the same 11,378,177-parameter model, objective, epoch count, and conditioning as the generator serving T2W and T2-FLAIR, trained on the same three subjects. We also compare against an L1-only T1W predecessor, affine/elastic augmentation, stand-alone 2.5D and residual-2.5D generators, paired Mixup, and output calibration. For server experiments only one component changes at a time; rows reusing a previous segmentation report image metrics only.

We do not compare against tuned external Pix2Pix, StarGAN, or diffusion implementations. A conditional flow-matching model was trained for T1W during development, but the only scored artifact we retain predates completion of its paired fine-tuning, so we decline to quote it; the absence of tuned external baselines is a limitation.

4 Results

4.1 Challenge Validation Results

Table 2 reports server results. Checkpoint averaging improves all three image metrics over the matched single-checkpoint submission, and T2-FLAIR fusion gives the largest single gain. The final row is the submitted container: mean SSIM 0.909452 with matched Dice 0.880905.

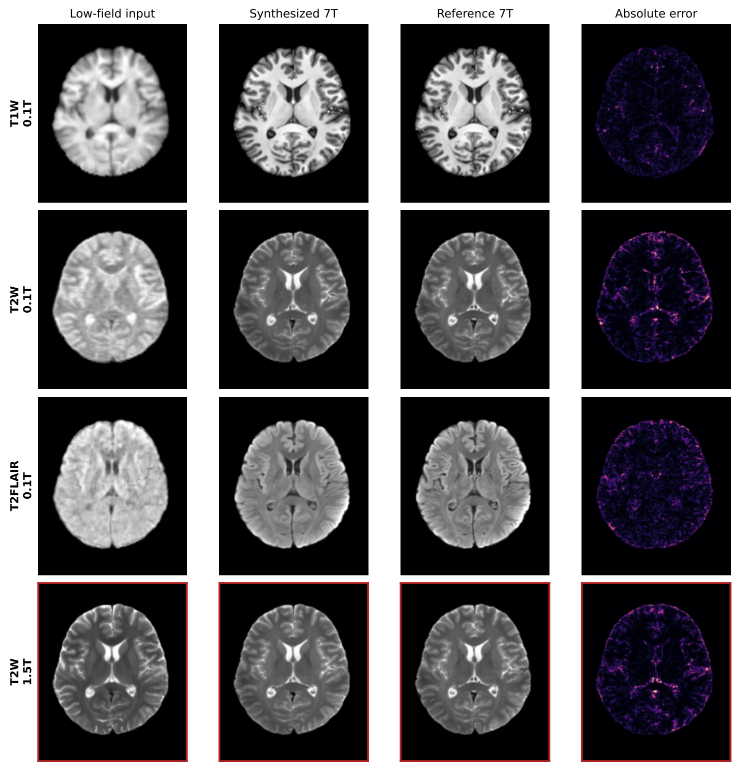

Fig. 2. Prospective subject 0006. Rows 1–3 show the three modality branches at 0.1T, the largest appearance gap in the challenge; row 4 shows T2W at 1.5T (red border), the weakest branch at low field, so the figure is not confined to a single source field. Images are percentile-normalized for display; error maps use the 7T intensity scale. Residual error concentrates at cortical boundaries and fine internal structures. This subject is training data, not a hidden validation target.

Two qualifications belong with this table. The final row was produced by regenerating the pipeline end to end rather than substituting one weight file, so small differences appear in all three modalities, including T2W, which the anchored stage cannot affect; the +0.0008 SSIM difference between the last two rows is thus attributable to the rebuild as a whole, not to anchoring alone, whose isolated evidence is Section 4.3. The configuration is also not uniformly best: its LPIPS (0.077601) is worse than the preceding row's 0.076841, and its mean Dice is below the matched single-checkpoint submission.

T2W is strongest (SSIM 0.9268, Dice 0.9128), while T1W and T2-FLAIR remain the bottlenecks. T2-FLAIR fusion raises SSIM from 0.8952 to 0.9001 and lowers nRMSE from 0.3277 to 0.3249, but lowers Dice from 0.8679 to 0.8473: it buys image fidelity at a structural cost, so a reader targeting segmentation should prefer the unfused branch.

Table 2. Challenge validation-server results (upper block) and the per-modality breakdown of the submitted container with matched segmentations (lower block). Dashes indicate that reconstruction-dependent labels were not regenerated for that historical image-only ablation.

Configuration	nRMSE↓	SSIM↑	LPIPS↓	Dice↑	Volume↑
Matched single-checkpoint submission	0.298081	0.906930	0.077610	**0.887631**	**0.891907**
+ T1W checkpoint averaging	0.297180	0.907342	0.077405	–	–
+ T2-FLAIR blend, $\alpha = 0.45$	0.295784	0.908641	**0.076841**	–	–
Submitted container (anchor, α=0.45)	**0.295467**	**0.909452**	0.077601	0.880905	0.886917
submitted, T1W	0.326192	0.901498	0.079743	0.882621	0.892379
submitted, T2W	0.235341	0.926802	0.063544	0.912752	0.917102
submitted, T2-FLAIR	0.324867	0.900057	0.089517	0.847343	0.851270

Table 3. Uniform field-conditioned supervised generator versus our T1W branch, on the three prospective subjects with 7T ground truth (12 volumes per row, scored slab).

	SSIM↑ by source field				Mean over 12	
T1W configuration	0.1T	1.5T	3T	5T	SSIM↑	nRMSE↓
Uniform supervised	0.9435	0.9590	0.9694	0.9657	0.959401	0.109037
Field-specific CUT, ep100	0.9820	0.9780	**0.9841**	**0.9815**	**0.981393**	0.066766
Shipped branch	**0.9825**	**0.9782**	0.9833	0.9809	0.981243	**0.066498**

4.2 Is One Supervised Model Enough?

A single field-conditioned supervised generator, identical in architecture and recipe to the one we use for T2W, is substantially worse on T1W (Table 3). It loses at every source field and on all twelve individual volumes, with 1.6× the error of the shipped branch, and the deficit is largest where the appearance gap is largest (−0.039 SSIM at 0.1T against −0.014 at 3T). Figure 2 shows the branches on the hardest setting.

With three paired subjects, supervised training alone cannot span the 0.1T-to-7T gap, whereas contrastive pretraining on 43–235 unpaired volumes per field can. Conversely, the same model is our strongest branch on T2W, where the paired relationship is stable. The split is thus an empirical finding, not a preference.

4.3 Generalization Ablations

Teacher anchoring improves both metrics in every held-out subject at every field: 12 of 12 subject–field combinations improve on both SSIM and nRMSE (one-sided sign test, $p = 2^{-12} = 2.4 \times 10^{-4}$), but the magnitude is of order 10^{-4}. Mean ΔSSIM is +0.000264, +0.000197, +0.000130, and +0.000016 for 0.1T, 1.5T, 3T, and 5T, with mean ΔnRMSE of −0.001151 to −0.000093 over the same fields. With $n = 3$ the Student-t 95% intervals on ΔSSIM exclude zero only at 1.5T, $[+0.000067, +0.000326]$. Both facts matter: the consistency of sign is evidence that restricting adaptation capacity does not hurt, while the magnitude does not support a claim of practical importance from three subjects. This is why we kept the 5T checkpoint ensemble.

Using a fixed epoch 6 across folds, mean 0.1T SSIM rises from 0.981968 to 0.982232 and nRMSE falls from 0.066838 to 0.065687. Alternatives fail on the same protocol. Against the frozen 2D baseline (SSIM 0.982012, nRMSE 0.066804), a stand-alone 2.5D generator is far worse despite paired fine-tuning (0.925011, 0.175642); a 2.5D residual corrector is neutral in SSIM (0.982041) while worsening nRMSE (0.068387); paired Mixup favors nRMSE (0.065663) but yields lower SSIM (0.982108) than anchoring alone. Limited paired subjects, not missing slice context, are the dominant constraint. Synchronized affine/elastic augmentation also fails to transfer, with server T1W SSIM falling from 0.900 to 0.892 and nRMSE rising to 0.365, and output calibration shifts mean LOO SSIM by only 1.8×10^{-5} while degrading subject 0009.

4.4 Masking, Selection Budget, and Cost

Background masking does not inflate the metrics. Aggregate metrics could in principle reflect background handling rather than intracranial synthesis. Scoring the submitted pipeline with masking enabled and disabled on all 36 prospective volumes changes mean nRMSE by +0.000521, SSIM by −0.000183, and LPIPS by +0.000119: masking is marginally *unfavourable*. Restricting metrics to anatomy confirms this — with nRMSE inside the source-derived mask and SSIM on a brain bounding box, nRMSE is essentially unchanged (0.0725 against 0.0718) and the modality ranking is preserved. The slab is already 47.7% foreground.

Selection budget. Our record contains at most 16 scored candidates for Task 1, at most 7 for T2-FLAIR fusion and 5 for T1W ensembles ("at most" because the count comes from archived bundles, not a pre-registered log). Only one component changed per submission, and all generalization claims were decided on subject-level LOO, never on the server. The weakest link is choosing $\alpha = 0.45$ over 0.40 on a 5×10^{-6} difference.

Cost. All 14 checkpoints fit on one GPU. On a single RTX 4090 at native resolution the pipeline peaks at 752 MiB of allocated GPU memory and needs 15.6 s per three-modality case (2.98–9.37 s per T1W volume by ensemble size, 2.90 s T2W, 6.51 s T2-FLAIR), so the 20-case test set needs about 5 minutes against a 20-hour limit. Training dominates: 20–30 GPU-hours of CUT pretraining per T1W model, and under 10 GPU-minutes for anchoring.

5 Discussion and Conclusion

The same supervised generator that is our strongest branch on T2W is markedly worse than contrastive translation on T1W at every source field, so what separates the cases is data regime rather than architecture. The teacher-anchor result is a consistency finding rather than a performance one: constrained adaptation never hurt any held-out subject, at 10^{-4} scale.

Several limitations bound these conclusions. Three prospective subjects cannot separate close checkpoints, and the gap between LOO SSIM near 0.98 and server SSIM near 0.91 indicates domain shift, so they poorly predict hidden-cohort performance. Some choices used server feedback, making our validation figures mildly optimistic. The supervised comparison uses our own model rather than tuned external baselines, and no external dataset, reader study, or lesion-preservation analysis was performed. Finally, synthesizing high-field appearance can produce fine structures the input does not support, which image-similarity metrics cannot detect, so these images should not be used for primary diagnostic reading. Overall the system reaches validation SSIM 0.9095 at 15.6 s per case, and its negative 2.5D, augmentation, and calibration results are equally informative; future work should prioritize larger subject-level cohorts and external baselines.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. vol. 27, pp. 2672–2680 (2014)
2. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 1125–1134 (2017). https://doi.org/10.1109/CVPR.2017.632
3. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232 (2017). https://doi.org/10.1109/ICCV.2017.244
4. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: ECCV 2020. LNCS, vol. 12354, pp. 319–345. Springer (2020). https://doi.org/10.1007/978-3-030-58545-7_19
5. Dar, S.U.H., Yurt, M., Karacan, L., Erdem, A., Erdem, E., Cukur, T.: Image synthesis in multi-contrast MRI with conditional generative adversarial networks. IEEE Trans. Med. Imaging **38**(10), 2375–2388 (2019). https://doi.org/10.1109/TMI.2019.2901750
6. Chartsias, A., Joyce, T., Dharmakumar, R., Tsaftaris, S.A.: Multimodal MR synthesis via modality-invariant latent representation. IEEE Trans. Med. Imaging **37**(3), 803–814 (2018). https://doi.org/10.1109/TMI.2017.2764326
7. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
8. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068
9. Billot, B., et al.: SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. Med. Image Anal. **86**, 102789 (2023). https://doi.org/10.1016/j.media.2023.102789

10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
11. Li, X., Grandvalet, Y., Davoine, F.: Explicit inductive bias for transfer learning with convolutional networks. In: ICML. PMLR, vol. 80, pp. 2825–2834 (2018)
12. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. vol. 30, pp. 1195–1204 (2017)