# Supplementary Material for ECaT

Meng Ye[1], Mikael Kanski[2], Dong Yang[3], Leon Axel[2], Dimitris Metaxas[1]
[1]Rutgers University, [2]New York University School of Medicine, [3]NVIDIA
{my389, dnm}@cs.rutgers.edu

## 1. More Detailed Results

### 1.1. Full Sequence Image-to-Image Translation and Multi-modal Image Registration Results

In Fig. 1, we show the image-to-image translation (I2I) and multi-modal image registration results on a full tagged MR (tMR) and untagged cine MR (cMR) image sequence pair. There are 24 frames in each sequence. As we discussed in the introduction section in the main text, in the early time frames of a tMR sequence, e.g., F0 and F1, the tagged blood obscures the boundary between myocardium wall and the blood pool, posing challenges on image registration and I2I. We thus trained a registration network, i.e., VoxelMorph, to warp the third frame in each tMR sequence to replace the first and second frames. Note that in all of our experiments, we did this pre-processing step on the tMR sequences for fair comparisons.

In Fig. 1, we use the checkerboard for the image alignment inspection. As indicated by the red arrows in column (g), before registration, cMR images are unaligned with tMR images. Our style reference-augmented I2I module can translate the tMR images into fake cMR images without any content distortion, as demonstrated in column (e), the good alignment between the fake cMR and corresponding tMR images. Our unsupervised cross-domain correspondence learning module can align features of each cMR image to be registered with those of the input tMR image and enforce sample-specific style consistency between the translated fake cMR image and the corresponding real cMR image. For example, the real cMR image style of frames F2, F3, F4 gradually changes from bright to dark, and our fake cMR images can successfully generate coherent image styles with their corresponding cMR style references. Similar examples could be found in frames F16, F17, F18, where both the real and fake cMR image style gradually changes from dark to bright. We qualitatively scrutinized all of our test dataset and found that all the fake cMR images possessed consistent image styles with their corresponding real cMR images to be registered. Note that, currently the temporal coherence in an image sequence is not exploited for a better I2I performance but we can achieve sample-specific style consistency for the fake images with our unsupervised cross-domain correspondence learning module. Since our I2I module can generate both content-preserving and style-coherent fake cMR images, the downstream image registration task can be greatly benefited, as demonstrated by the good alignment between the registered cMR and tMR images shown in column (f).

### 1.2. Learned Cross-domain Correspondences

In Fig. 2, we show more detailed comparison of learned cross-domain correspondences between our model and A3.

(1) The query points in patch 1 of tMR are **inside** the cardiac wall, but A3 predicts correspondences outside the wall in the style reference (SR), i.e., cMR.

(2) The query points in patch 2 and 4 of tMR are **on the boundary** of the cardiac wall, but A3 predicts correspondences outside and far away from the wall in the SR.

(3) The query points in patch 3 of tMR are slightly **outside** the cardiac wall, but A3 predicts correspondences inside the wall in the SR.

Our model, however, can learn plausible cross-domain correspondences without supervision for all kinds of query points. Our I2I module thus generates the sample-specific style consistent fake image with the style reference image.

### 1.3. Visualization of Deformation Field

In Fig. 3, we visualize the predicted deformation fields by different methods. As we discussed in the introduction section in the main text, large deformation between modalities is one of the challenges for multi-modal medical image registration. We decompose large deformation between the tMR and cMR image pair as global affine and local non-rigid deformation components. By the design of a shared TransUnet for efficient feature embedding, our registration network can predict both affine and non-rigid deformations simultaneously. As shown in Fig. 3 (b), our predicted deformation fields are the composition results of affine and non-rigid deformation components. In this way, while the affine component first deforms the moving image coarsely to eliminate possible linear and large spatial deformation, the non-rigid component then deforms the coarsely aligned moving image in a finer scale to further eliminate non-linear and small misalignment beyond the coverage of the affine component. Our method generates smooth, invertible deformation fields while capturing large deformations between different modality images. Multi-resolution deformation decomposition is another efficient way for the estimation of large deformations, as used in the MIND baseline method (with a 3-level image pyramid). However, by comparing Fig. 3 (b) and (c), we note that, without the estimation of possible affine deformation component, multi-resolution method is less efficient than our method and struggles to estimate large deformations, as shown in the 3 chamber and 4 chamber views in Fig. 3.

We note that, accurate and efficient joint estimation of global affine and local non-rigid deformations is non-trivial. While there exist methods of using a shrinking network for affine deformation estimation and a shrinking-expanding network for non-rigid deformation estimation, few work tries to share the parameters for the two networks. In [1], an encoder-decoder based U-Net is employed to estimate affine and non-rigid deformations. More specifically, the encoder together with fully connected layers output the affine deformation parameters and the decoder outputs non-rigid deformation parameters. Then the two kinds of deformations are composed as the hybrid deformation to warp the moving image towards the fixed image. In this way, the two kinds of deformations share the encoder of the U-Net. Note that, unlike cascaded methods which warp the moving image multiple times, this method only warp the moving image once with the estimated hybrid deformation field, which is efficient but brings limitations. Firstly, it brings instability for network training because the encoder is susceptible to predict large affine deformations. While this limitation could be mitigated by initializing the network training with only the non-rigid deformation prediction, the method is designed to predict misalignment between binary segmentation masks. However, the registration between images needs to optimize the network under the gray-scale intensity space which is more complex than the binary space. Thus, this method cannot generalize well from registering binary masks to gray-scale images. We re-implemented this method and trained S1 and S2 with the same generator as our model which was trained in the first stage. From the results in Table 1, without cascading the deformations, both S1 (based on U-Net) and S2 (based on TransUnet) are inferior to our method. We see the method in [1] fails to accurately estimate the affine deformations between gray-scale images.

We further note that, to our best knowledge, no previous work uses a shrinking-expanding network for affine deformation estimation and we are the first to use a shared shrinking-expanding subnetwork for the two stages of deformation estimation. Our method not only reduces network parameters greatly but also enables joint estimation of the two kinds of deformation components compactly. While the parameter reduction can avoid overfitting, joint estimation of affine and non-rigid deformations results in a global optimum of the final composed deformation field. Our method thus outperforms all the baseline multi-modal medical image registration methods by a significant margin.

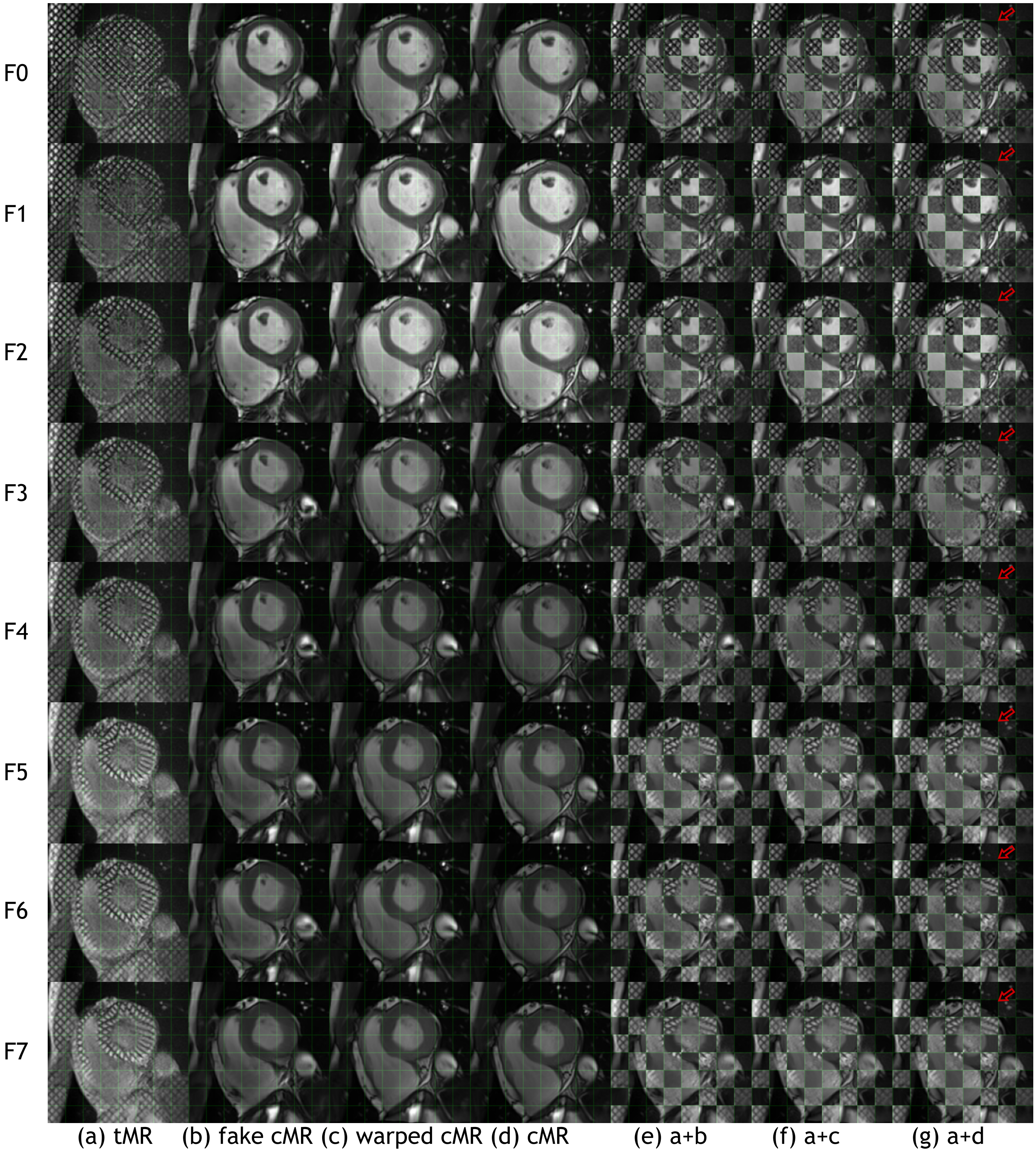
| Model | Cascading | Shared Parts | Embedding Layer | Dice (%) ↑ | | |
|---|---|---|---|---|---|---|
| | | | | Affine | Non-rigid | Composed |
| S1 [1] | | Encoder | CNN | $65.6 \pm 15.9$ | $75.7 \pm 13.3$ | $75.8 \pm 13.2$ |
| S2 | | Encoder | ViT | $66.8 \pm 16.1$ | $\mathbf{76.1 \pm 13.0}$ | $76.2 \pm 13.2$ |
| Ours | ✓ | Encoder+Decoder | ViT | $\mathbf{69.6 \pm 14.5}$ | $75.4 \pm 13.5$ | $\mathbf{77.4 \pm 11.9}$ |

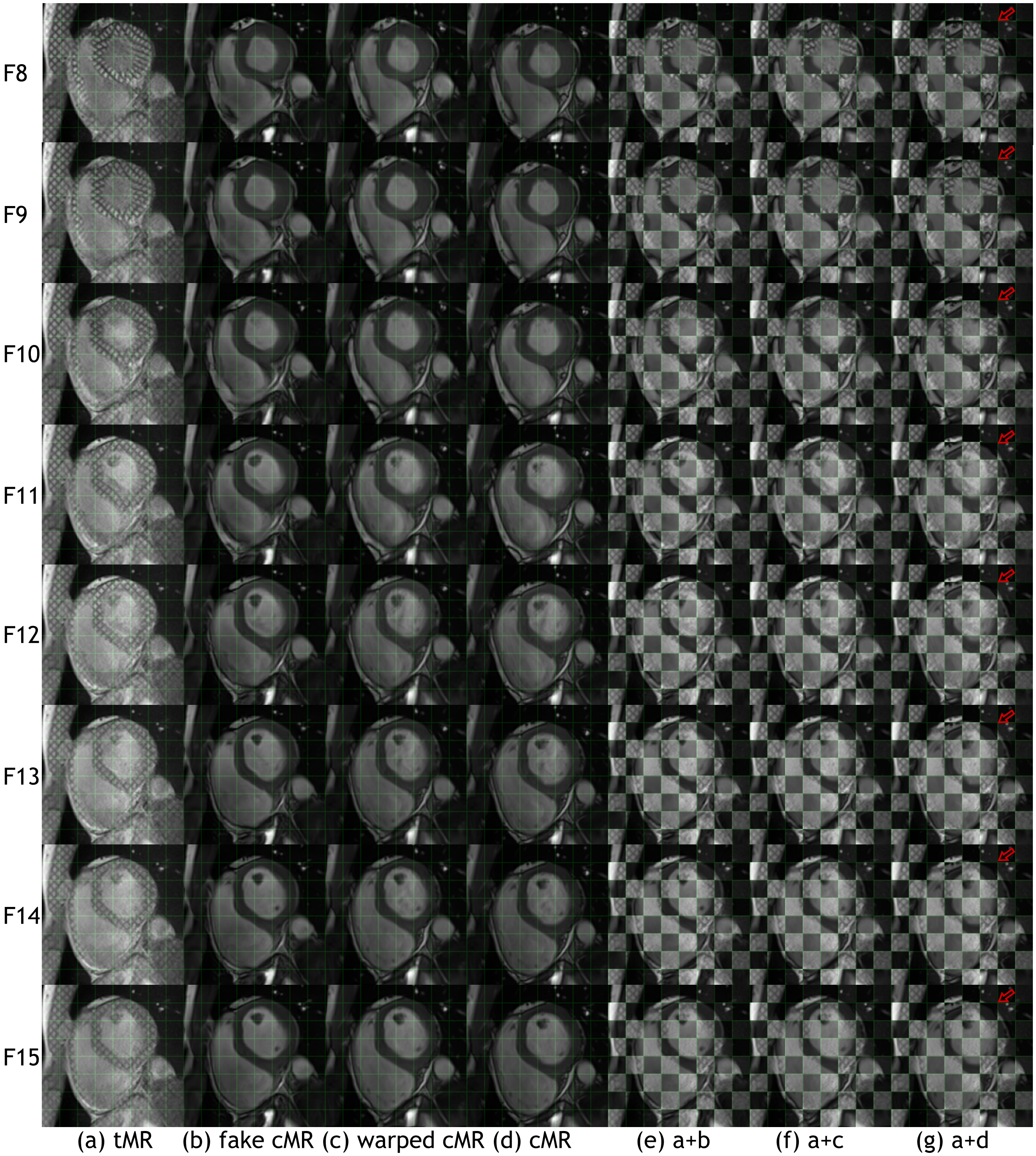
Table 1. Comparison of joint affine and non-rigid deformation estimation performance between [1] and ours.

Lastly, we clarify why the network parameters of our model are less than those of C2 in Table 5 in the main text. We use a 5-level TransUnet for both models as the shared feature embedding subnetwork for the two stages of deformation estimation. The feature embedding dimensions in each level of the encoder are $2, 16, 32, 32, 32$, while in each level of the decoder are $32, 32, 32, 32, 16$. For C2, the input feature embedding for the affine deformation estimation head comes from the last layer of the encoder, which has a dimension of 32; for our model, it comes from the last layer of the decoder, which has a dimension of 16. The larger input feature embedding dimension for C2 requires more fully connected neurons in the affine deformation estimation head, thus resulting in slightly larger network parameter size.

## References

[1] Matthew Sinclair, Andreas Schuh, Karl Hahn, Kersten Petersen, Ying Bai, James Batten, Michiel Schaap, and Ben Glocker. Atlas-istn: Joint segmentation, registration and atlas construction with image-and-spatial transformer networks. *Medical Image Analysis*, 78:102383, 2022. 2


F0
F1
F2
F3
F4
F5
F6
F7
(a) tMR
(b) fake cMR
(c) warped cMR
(d) cMR
(e) a+b
(f) a+c
(g) a+d


F8
F9
F10
F11
F12
F13
F14
F15
(a) tMR
(b) fake cMR
(c) warped cMR
(d) cMR
(e) a+b
(f) a+c
(g) a+d

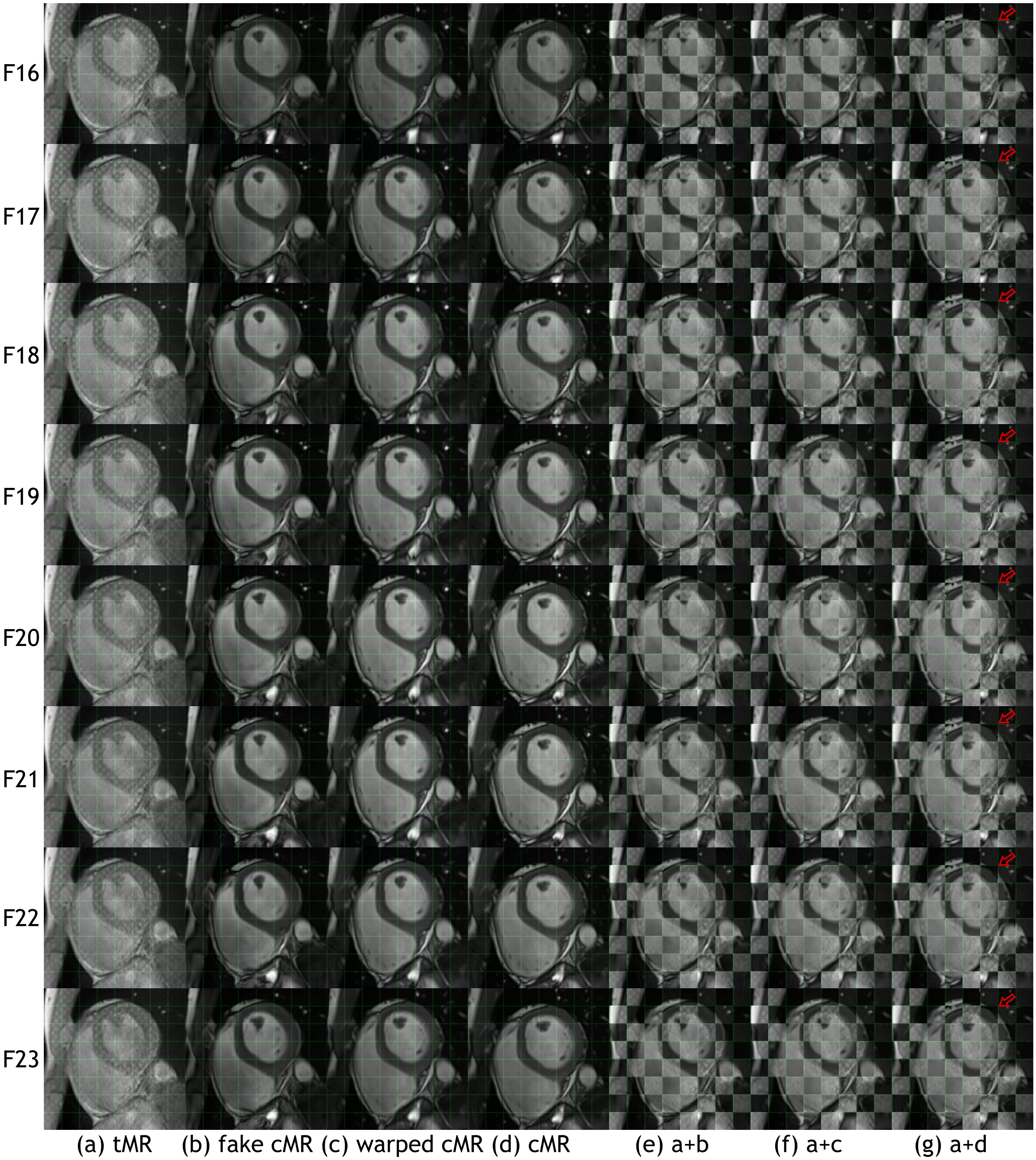


Figure 1. Image-to-image translation and multi-modal image registration results shown on a full tMR and cMR image sequence pair (best viewed zoomed in). (a) tMR; (b) fake cMR; (c) registered cMR; (d) cMR; (e) checkerboard of (a) and (b); (f) checkerboard of (a) and (c); (g) checkerboard of (a) and (d). ‘F’ means ‘frame’. Red arrows highlight unaligned areas between tMR and cMR.

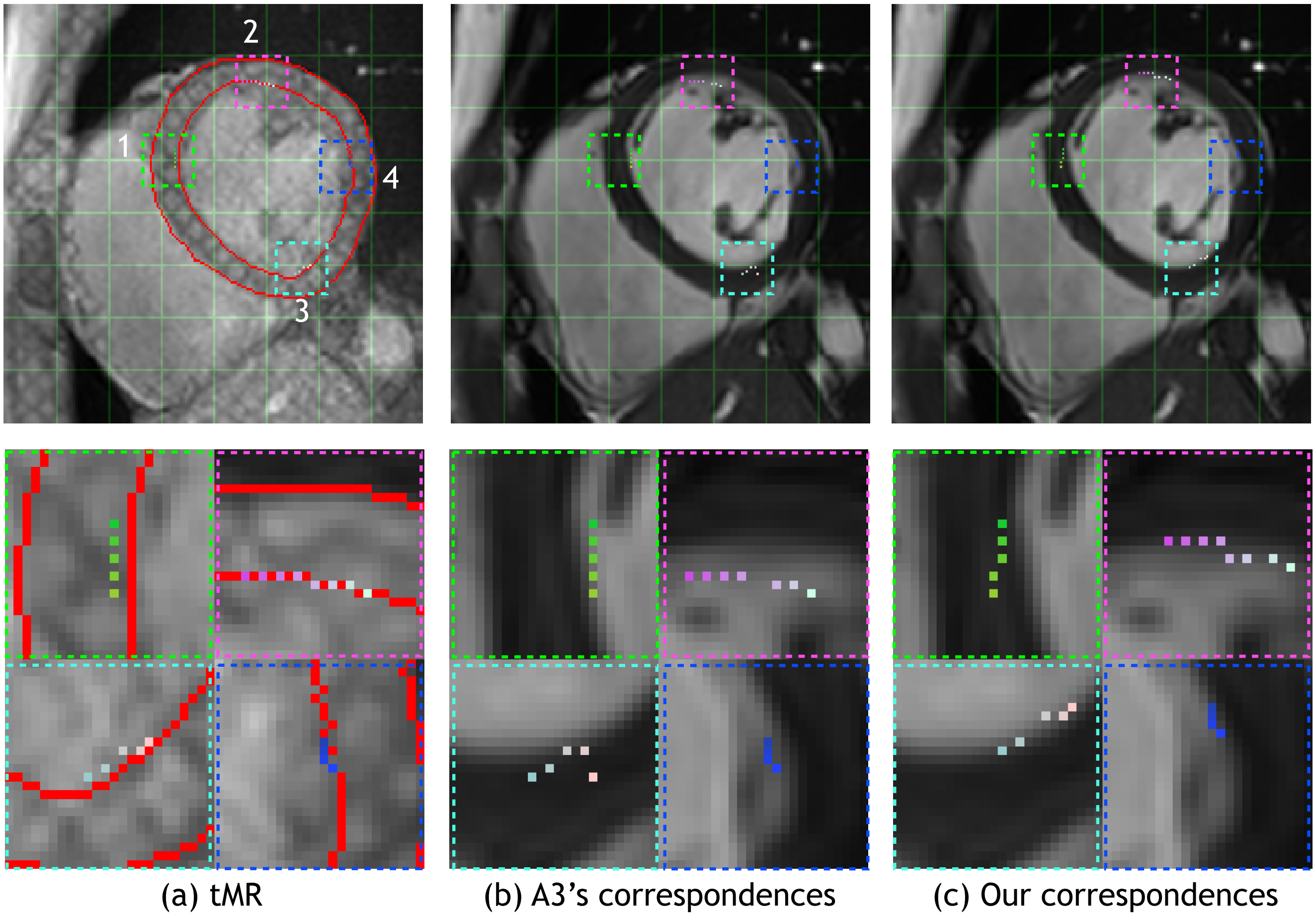


Figure 2. Learned cross-domain correspondences (color-coded) between a tMR and cMR (the style reference) image pair by different I2I models. The cMR image is unaligned with the tMR image. For the query points in tMR (a), our model (c) can predict plausible cross-domain correspondences in the style reference cMR in an unsupervised fashion. Due to the lack of effective supervision and without a content-preserving loss, A3 (b) cannot learn reasonable cross-domain correspondences. First row shows tMR image and query points (left); style reference cMR image and learned correspondences by A3 (middle); style reference cMR image and learned correspondences by our model (right). Second row shows zoomed in patches. Red contour shows the ground truth myocardium wall on tMR. Note the many-to-one mapping (8 query points with 7 correspondences) in patch 2 for A3.

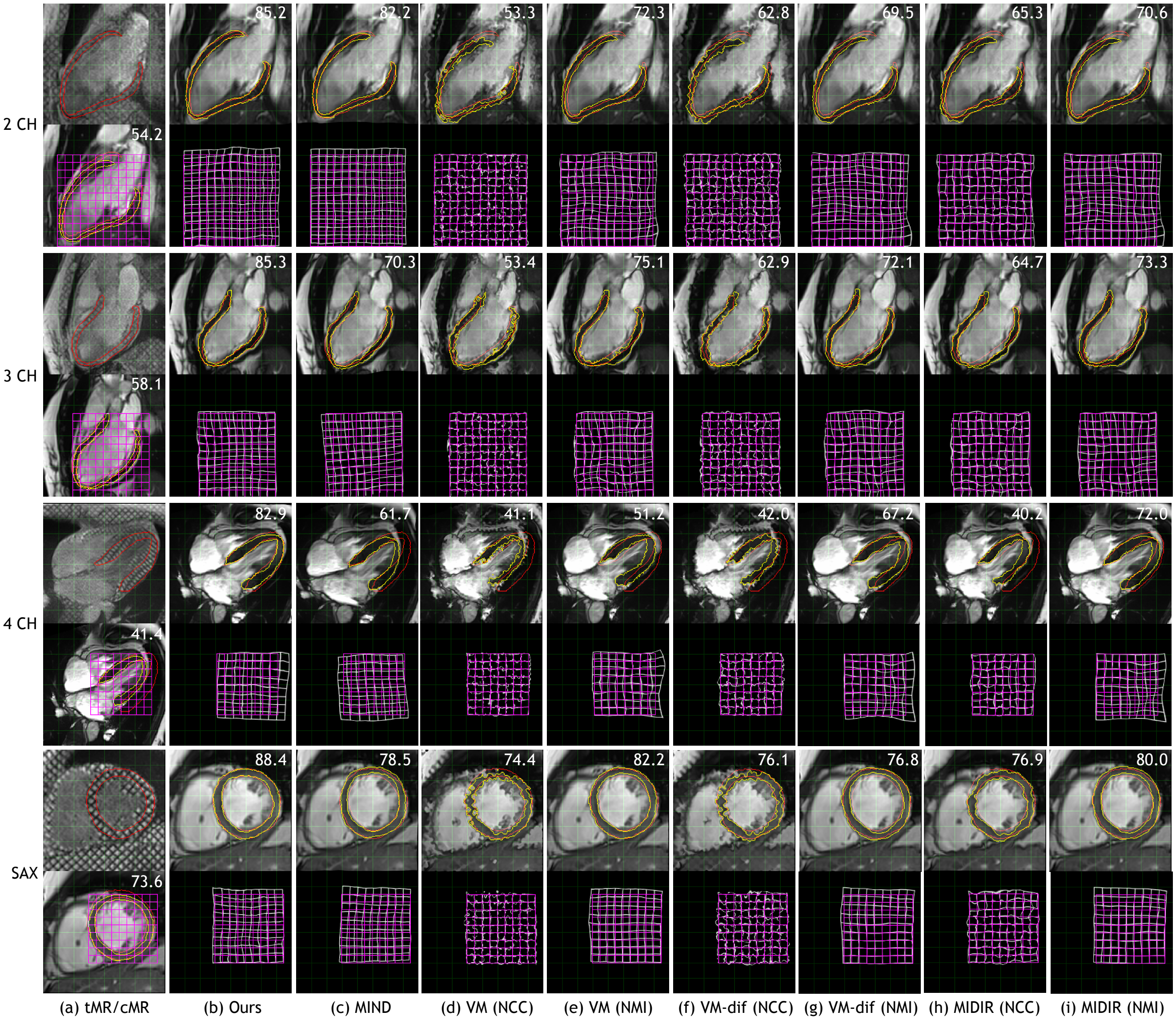


Figure 3. Visualization of the deformation field. We show four examples from 2/3/4 chamber (CH) view and short axis (SAX) view. For each view, we show the tMR/cMR image in (a), the registered cMR by different methods and corresponding deformation field (white grid) in (b)∼(i). Red/yellow contour shows the ground truth/warped myocardium wall on tMR/(warped) cMR. The right top of each (warped) cMR shows the Dice score. Pink grid shows the identity deformation field near the left ventricle area. We overlap the identity deformation field with the predicted field for convenient comparison.