Supplemental Materials of "RFNet: Unsupervised Network for Mutually Reinforcing Multi-modal Image Registration and Fusion"

Han Xu[1], **Jiayi Ma**[1*], **Jiteng Yuan**[1], **Zhuliang Le**[1], **and Wei Liu**[2]
[1] Wuhan University, China [2] Tencent Data Platform, China
{xu_han,yuanjiteng,lezhuliang}@whu.edu.cn, jyma2010@gmail.com, wl2223@columbia.edu

1. Illustration of Existing Sparse Descriptors

Transformation-based image registration methods transfer images into a common space to exhibit better consistency. Representative methods include entropy image entropy image (EI) [6], dense adaptive self-correlation (DASC) local descriptor [4, 5], structure consistency boosting (SCB) transform [1], *etc*. The transformed results of them are usually sparse, as shown in Fig. 1, which are not conductive to the network convergence.

2. Details of Network Architecture

The details of network architecture are reported in Tabs. 1, 2 and Fig. 2. For image translation, we apply the instance normalization as it normalizes feature statistics. Given an input batch $u \in \mathbb{R}^{N\times H\times W\times C}$ (N, H, W and C: batch size, height, width and channel), the instance normalized u is calculated as:

$$\mathrm{IN}(u) = \gamma(\frac{u-\bar{u}}{\sigma}) + \beta, \tag{1}$$

where $\bar{u}$ and $\sigma \in \mathbb{R}^{N\times C}$ are the mean and standard deviation across spatial dimensions. γ and β are the scale and shift parameters.

For image registration, we use the deformable convolution to adapt to the deformation in unregistered images. More concretely, the process of traditional convolution is mathematically defined as:

$$y(p) = \sum_{p_n\in\mathcal{R}} w(p_n)\cdot x(p+p_n) + b, \tag{2}$$

where $x(p)$ is the value of input x at pixel p. y is the output feature map. w and b are the weights and bias, respectively. Take the 3×3 convolution kernel as an example, $\mathcal{R} = \{(-1,-1),(-1,0),\cdots,(1,1)\}$. It can be seen that the receptive field $\mathcal{R}$ is a regular region. By comparison, the deformable convolution is represented as:

$$y(p) = \sum_{p_n\in\mathcal{R}} w(p_n)\cdot x(p+p_n+\triangle p_n) + b, \tag{3}$$

where $\triangle p_n$ denotes the offsets learned and generated from additional convolution layers. It includes both horizontal and vertical offsets. Considering that $\triangle p_n$ may be a decimal, $x(p+p_n+\triangle p_n)$ is obtained by bilinear interpolation.

Then, through the global average pooling (GAP) layer across spatial dimensions, the feature maps are mapped into a 128-dimensional vector. Compared with fully connected layer, the GAP layer refers to any size of input. Thus, there is no limit on the size of original multi-modal images. Finally, we feed the 128-dimensional vector into a fully connected layer to generate the final affine parameters.

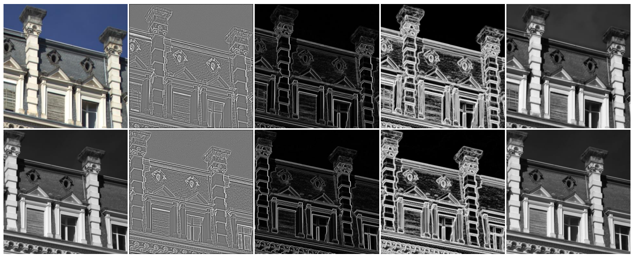

Figure 1. Descriptors of existing methods for image registration and the translated results of the proposed RFNet. From left to right: VIS/NIR images, descriptors of WLD [2], NTG [3] and SCB [1], and translated NIR/real NIR obtained in our method.

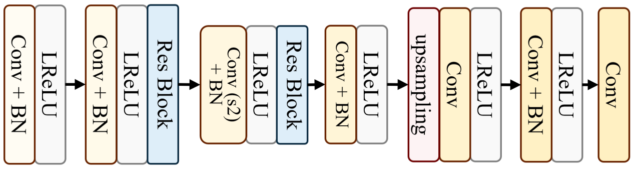

Figure 2. Pipeline of the deformation block ("s2": stride set to 2. "Res Block": residual block. "BN": batch normalization).

[*] Corresponding author

Table 1. Network architecture of TransNet. Conv(Nα, Kβ, Sγ) means a convolution layer with α filters, where kernel size is $\beta \times \beta$ and stride is γ. H is the channel of generated images. LReLU is the LeakyReLU with slope set as 0.2. Deconv is a deconvolution layer. IN denotes the instance normalization.

TransNet - Encoder	Shared
1 Conv (N16, K7, S1) - ReLU	✗
2 Conv (N64, K3, S1) - IN - ReLU	✗
3 Conv (N128, K3, S2) - IN - ReLU	✗
4 Residual Block (N128)	✗
5 Residual Block (N128)	✗
6 Residual Block (N128)	✓
TransNet - Decoder	
1 Residual Block (N128)	✓
2 Residual Block (N128)	✗
3 Residual Block (N128)	✗
4 Residual Block (N128)	✗
5 Deconv (N64, K3, S1) - IN - LReLU	✗
6 Conv (NH, K7, S1) - tanh	✗

Table 2. Network architecture of AffineNet. GAP means the global average pooling and FC denotes the fully connected layer. DeforConv is the deformable convolution layer with kernel size set as 3×3 and stride set as 1. "Max" denotes maxpooling.

AffineNet
1 Conv(N16, K7, S1)-LReLU-DeforConv(N16)-LReLU
2 Conv(N32, K7, S2)-LReLU-DeforConv(N32)-LReLU
3 Conv(N64, K7, S2)-LReLU-DeforConv(N64)-LReLU
4 Conv(N64, K7, S2)-LReLU-DeforConv(N64)-LReLU-Max
5 Conv(N128, K7, S1)-LReLU-DeforConv(N128)-LReLU-Max
6 GAP - FC (N6)

3. Illustration of Landmark

To validate the registration accuracy, we manually build 5 pairs of point landmarks in each image pair, as shown in Fig. 3. The source points are represented as $\{(x_1^s, y_1^s), \cdots, (x_5^s, y_5^s)\}$ while the target ones are defined as $\{(x_1^t, y_1^t), \cdots, (x_5^t, y_5^t)\}$. In the registered image, the source points are registered into transformed points $\{(x_1^r, y_1^r), \cdots, (x_5^r, y_5^r)\}$. Then, the metrics including root mean square error (RMSE), max square error (MAE) and median square error (MEE) are measured based on the distances between $\{(x_1^r, y_1^r), \cdots, (x_5^r, y_5^r)\}$ and $\{(x_1^t, y_1^t), \cdots, (x_5^t, y_5^t)\}$.

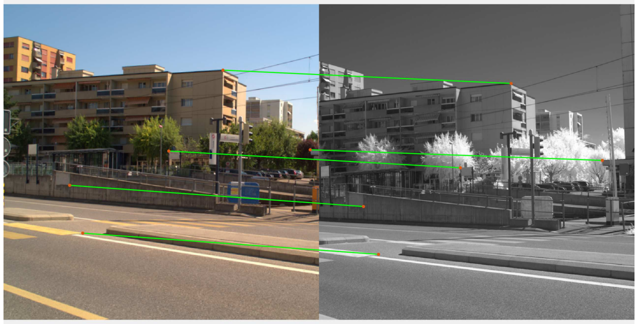

Figure 3. Illustration of five points of landmarks in a pair of multi-modal images (left: VIS image, right: NIR image).

4. Evaluation Metrics of Image Fusion

This section introduces the definitions of metrics evaluating the fusion performance. These metrics include average gradient (AG), entropy (EN), standard deviation (STD) and peak signal-to-noise ratio (PSNR). Denoting the source images as I_a and I_b and the fused image is represented as I_f, the mathematical definitions of these metrics are as follows.

- Average gradient (AG)
  AG quantifies the gradient information of I_f and represents the details and textures in I_f. A large AG indicates that I_f contains more content information and texture details and thus exhibits a better fusion performance. Specifically, it is defined as:

$$AG = \frac{1}{MN} \sum_{2}^{M} \sum_{2}^{N} \sqrt{\frac{\nabla I_{f_x}^2(i,j) + \nabla I_{f_y}^2(i,j)}{2}}, \tag{4}$$

  where $\nabla I_{f_x}(i,j) = I_f(i,j) - I_f(i-1,j)$ and $\nabla I_{f_y}(i,j) = I_f(i,j) - I_f(i,j-1)$. M and N are the width and height of I_f.

- Entropy (EN)
  EN measures the richness of information in a fused image on the basis of information theory. A larger EN always means that more information is contained in the fused image and the performance of the fusion method is better. EN is defined as follows:

$$EN = -\sum_{l=0}^{L-1} p_l \log_2 p_l, \tag{5}$$

  where L denotes the number of gray levels and p_l is the normalized histogram of the corresponding gray level in the fused image.

- Standard deviation (STD)
  STD reflects the distribution and contrast of the fused

image form a statistical view, which is mathematically defined as follows:

$$STD = \sqrt{\sum_{1}^{M}\sum_{1}^{N}(I_f(i,j)-\mu)^2}, \qquad (6)$$

where μ denotes the mean value of the fused image I_f. A fused image with a large STD indicate that it has a high contrast.

- Peak signal-to-noise ratio (PSNR)
  PSNR measures the ratio of peak value power and noise power of the fused image. It reflects the distortion during the fusion process. Mathematically, it is defined as follows:

$$PSNR = 10\log_{10}\frac{r^2}{\frac{\|I_f-I_a\|_2+\|I_f-I_b\|_2}{2}}, \qquad (7)$$

  where r is the peak value of I_f. A large PSNR demonstrates that I_f is similar to I_a and I_f and the fusion process produces little distortion.

References

[1] Si-Yuan Cao, Hui-Liang Shen, Shu-Jie Chen, and Chunguang Li. Boosting structure consistency for multispectral and multimodal image registration. *IEEE Transactions on Image Processing*, 29:5147–5162, 2020. 1

[2] Jie Chen, Shiguang Shan, Chu He, Guoying Zhao, Matti Pietikäinen, Xilin Chen, and Wen Gao. Wld: A robust local image descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1705–1720, 2009. 1

[3] Shu-Jie Chen, Hui-Liang Shen, Chunguang Li, and John H Xin. Normalized total gradient: a new measure for multispectral image registration. *IEEE Transactions on Image Processing*, 27(3):1297–1310, 2017. 1

[4] Seungryong Kim, Dongbo Min, Bumsub Ham, Minh N Do, and Kwanghoon Sohn. Dasc: Robust dense descriptor for multi-modal and multi-spectral correspondence estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1712–1729, 2017. 1

[5] Seungryong Kim, Dongbo Min, Bumsub Ham, Seungchul Ryu, Minh N Do, and Kwanghoon Sohn. Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2103–2112, 2015. 1

[6] Christian Wachinger and Nassir Navab. Structural image representation for image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition-Workshops*, pages 23–30, 2010. 1