DDFM: Denoising Diffusion Model for Multi-Modality Image Fusion
SUPPLEMENTARY MATERIALS

Zixiang Zhao[1,2] Haowen Bai[1] Yuanzhi Zhu[2] Jiangshe Zhang[1*] Shuang Xu[3]
Yulun Zhang[2] Kai Zhang[2] Deyu Meng[1,5] Radu Timofte[2,4] Luc Van Gool[2]

[1]Xi'an Jiaotong University [2]Computer Vision Lab, ETH Zürich
[3]Northwestern Polytechnical University [4]University of Würzburg
[5]Macau University of Science and Technology

zixiangzhao@stu.xjtu.edu.cn, jszhang@mail.xjtu.edu.cn

Abstract

In this document, we provide the additional supplementary information for the paper "DDFM: Denoising Diffusion Model for Multi-Modality Image Fusion". This file contains:
(I) The detail algorithm architecture for DDIM Sampling which is mentioned in Sec. 2.1.
(II) The detailed derivations for our DDFM in Sec. 3.
(III) Detailed illustration to the training&testing datasets in Sec. 4.1.
(IV) Detailed introduction for the selection and analysis of hyperparameters in Sec. 4.1.
(V) More qualitative comparison fusion results in Sec. 4.2.

1. Detailed introduction for DDPM Sampling

We show the DDPM sampling algorithm for the DDIM fashion [3] in Algorithm 1. For comparison, we show the simplified version of our DDFM in Algorithm 2. Obviously, the estimate $\tilde{\boldsymbol{f}}_{0|t}$ is used to predict $\boldsymbol{f}_{t-1}$ in vanilla DDPM (*line 8*). However, in our DDFM, $\tilde{\boldsymbol{f}}_{0|t}$ is utilized as the initial input for the EM module, and we rectify $\tilde{\boldsymbol{f}}_{0|t}$ to $\hat{\boldsymbol{f}}_{0|t}$ to meet the likelihood in the EM module. Then, $\hat{\boldsymbol{f}}_{0|t}$ is used to predict $\boldsymbol{f}_{t-1}$.

[*] Corresponding author.

Algorithm 1 Vanilla DDIM

```
Input:
    $T$, $\{\tilde{\sigma}_t\}_{t=1}^{T}$
Output:
    Generated image $\boldsymbol{f}_0$.
$\boldsymbol{f}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T-1$ to $0$ do
    % DDPM Part 1: Obtain $\tilde{\boldsymbol{f}}_{0|t}$
    $\hat{\boldsymbol{s}} \leftarrow \boldsymbol{s}_\theta\left(\boldsymbol{f}_t, t\right)$
    $\tilde{\boldsymbol{f}}_{0|t} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{f}_t + (1-\bar{\alpha}_t)\,\hat{\boldsymbol{s}}\right)$
    % DDPM Part 2: Estimate $\boldsymbol{f}_{t-1}$
    $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\boldsymbol{f}_{t-1} \leftarrow \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{f}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\tilde{\boldsymbol{f}}_{0|t} + \tilde{\sigma}_t \boldsymbol{z}$
end for
```

Algorithm 2 Our DDFM (to compare with Vanilla DDIM)

```
Input:
    Infrared image $\boldsymbol{i}$, Visible image $\boldsymbol{v}$, $T$, $\{\tilde{\sigma}_t\}_{t=1}^{T}$
Output:
    Fused image $\boldsymbol{f}_0$.
$\boldsymbol{f}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T-1$ to $0$ do
    % DDPM Part 1: Obtain $\tilde{\boldsymbol{f}}_{0|t}$
    $\hat{\boldsymbol{s}} \leftarrow \boldsymbol{s}_\theta\left(\boldsymbol{f}_t, t\right)$
    $\tilde{\boldsymbol{f}}_{0|t} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{f}_t + (1-\bar{\alpha}_t)\,\hat{\boldsymbol{s}}\right)$
    % E-step: Update latent variables
    % M-step: Obtain $\hat{\boldsymbol{f}}_{0|t}$ via Likelihood Rectification
    % DDPM Part 2: Estimate $\boldsymbol{f}_{t-1}$
    $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\boldsymbol{f}_{t-1} \leftarrow \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{f}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\boldsymbol{f}}_{0|t} + \tilde{\sigma}_t \boldsymbol{z}$
end for
```

2. Detailed derivations of some equations in Sec. 3

Eq. (11):

$$\begin{aligned}
&\mathcal{LAP}(\xi;\mu,\sqrt{b/2})\\
=&\frac{1}{2}\sqrt{\frac{2}{b}}\exp\left(-\sqrt{\frac{2}{b}}|\xi-\mu|\right)\\
=&\int_0^\infty \frac{1}{\sqrt{2\pi a}}\exp\left(-\frac{(\xi-\mu)^2}{2a}\right)\frac{1}{b}\exp\left(-\frac{a}{b}\right)da\\
=&\int_0^\infty \mathcal{N}(\xi;\mu,a)\mathcal{EXP}(a;b)da
\end{aligned} \tag{1}$$

Eq. (17):

$$\begin{aligned}
&\log p(\tilde{m}_{ij};y_{ij},x_{ij})\\
=&\log p(y_{ij};x_{ij},a_{ij})+\log p(\tilde{m}_{ij})\\
=&-\frac{\log a_{ij}}{2}-\frac{(y_{ij}-x_{ij})^2}{2a_{ij}}-2\log\tilde{m}_{ij}-\frac{1}{\gamma\tilde{m}_{ij}}+\text{constant}\\
=&\frac{\log\tilde{m}_{ij}}{2}-\frac{\tilde{m}_{ij}(y_{ij}-x_{ij})^2}{2}-2\log\tilde{m}_{ij}-\frac{1}{\gamma\tilde{m}_{ij}}+\text{constant}\\
=&-\frac{3}{2}\log\tilde{m}_{ij}-\frac{\tilde{m}_{ij}(y_{ij}-x_{ij})^2}{2}-\frac{1}{\gamma\tilde{m}_{ij}}+\text{constant}.
\end{aligned} \tag{2}$$

Eq. (19):

$$\begin{aligned}
&\log p(\tilde{n}_{ij};x_{ij})\\
=&\log p(x_{ij};b_{ij})+\log p(\tilde{n}_{ij})\\
=&-\frac{\log b_{ij}}{2}-\frac{x_{ij}^2}{2b_{ij}}-2\log\tilde{n}_{ij}-\frac{1}{\rho\tilde{n}_{ij}}+\text{constant}\\
=&-\frac{3}{2}\log\tilde{n}_{ij}-\frac{\tilde{n}_{ij}x_{ij}^2}{2}-\frac{1}{\rho\tilde{n}_{ij}}+\text{constant}.
\end{aligned} \tag{3}$$

Eq. (25):
Step 1: Convolution theorem.

The convolution theorem states that the convolution operation in the spatial domain is equivalent to element-wise multiplication in the frequency domain. The convolution theorem can be expressed as:

$$\text{fft}\{A * x\} = \text{fft}\{A\} \odot \text{fft}\{x\},$$

where $*$ is the convolution operation.

Step 2: Solve for the Fourier transform of the solution in the frequency domain.

Now, let's substitute the convolution theorem into the optimization problem. In Eq. (24), we have:

$$\min_{\boldsymbol{k}} \mathcal{L}_{\boldsymbol{k}} = ||\boldsymbol{k} - \boldsymbol{x}||_2^2 + ||\boldsymbol{u} - \nabla \boldsymbol{k}||_2^2.$$

Then in frequency domain:

$$\min_{\text{fft}\{\boldsymbol{k}\}} \mathcal{L}_{\text{fft}\{\boldsymbol{k}\}} = \|\text{fft}\{\boldsymbol{k}\} - \text{fft}\{\boldsymbol{x}\}\|_2^2 + \|\text{fft}\{\nabla\} \odot \text{fft}\{\boldsymbol{k}\} - \text{fft}\{\boldsymbol{u}\}\|_2^2$$

Since we are working in the frequency domain, the problem becomes a simple least squares problem:

$$\text{fft}\{\hat{\boldsymbol{k}}\} = \arg\min_{\text{fft}\{\boldsymbol{k}\}} \|\text{fft}\{\boldsymbol{k}\} - \text{fft}\{\boldsymbol{x}\}\|_2^2 + \|\text{fft}\{\nabla\} \odot \text{fft}\{\boldsymbol{k}\} - \text{fft}\{\boldsymbol{u}\}\|_2^2$$

To solve for $\text{fft}\{\hat{\boldsymbol{k}}\}$, we can set the derivative with respect to $\text{fft}\{\boldsymbol{k}\}$ to zero:

$$\frac{\partial}{\partial \text{fft}\{\boldsymbol{k}\}} \left(\|\text{fft}\{\boldsymbol{k}\} - \text{fft}\{\boldsymbol{x}\}\|_2^2 + \|\text{fft}\{\nabla\} \odot \text{fft}\{\boldsymbol{k}\} - \text{fft}\{\boldsymbol{u}\}\|_2^2 \right) = 0$$

Simplifying and setting the derivative to zero gives:

$$\overline{\text{fft}\{\nabla\}} \odot (\text{fft}\{\nabla\} \odot \text{fft}\{\boldsymbol{k}\} - \text{fft}\{\boldsymbol{u}\}) + (\text{fft}\{\boldsymbol{k}\} - \text{fft}\{\boldsymbol{x}\}) = 0$$

Where $\overline{\text{fft}\{\nabla\}}$ is the complex conjugation of $\text{fft}\{\nabla\}$.

Solving for $\text{fft}\{\boldsymbol{k}\}$, we get:

$$\text{fft}\{\boldsymbol{k}\} = \frac{\text{fft}(\boldsymbol{x}) + \overline{\text{fft}(\nabla)} \odot \text{fft}(\boldsymbol{u})}{1 + \overline{\text{fft}(\nabla)} \odot \text{fft}(\nabla)}$$

Step 3: Inverse transform to get the solution in the spatial domain.

Now that we have the solution in the frequency domain, we can take the inverse Fourier transform to obtain the solution in the spatial domain:

$$\boldsymbol{k} = \text{ifft}\left\{ \frac{\text{fft}(\boldsymbol{x}) + \overline{\text{fft}(\nabla)} \odot \text{fft}(\boldsymbol{u})}{1 + \overline{\text{fft}(\nabla)} \odot \text{fft}(\nabla)} \right\}$$

This is the final expression for the solution $\boldsymbol{k}$ in the spatial domain.

3. Detailed introduction to datasets

We adopt widely-used benchmarks MSRS [4], RoadScene [6], M^{3}FD [2] and TNO [5] for *Infrared-Visible image Fusion* (IVF), and Harvard Medical Image Dataset [1] for *Medical Image Fusion* (MIF), respectively.

- MSRS dataset[1]: 50 pairs for IVF testing.
- RoadScene dataset[2]: 50 pairs for IVF testing.
- TNO dataset[3]: 25 pairs for IVF testing.
- M^{3}FD dataset[4]: 50 pairs for IVF testing.
- Harvard Medical Image dataset[5]: 25 pairs for MIF testing.

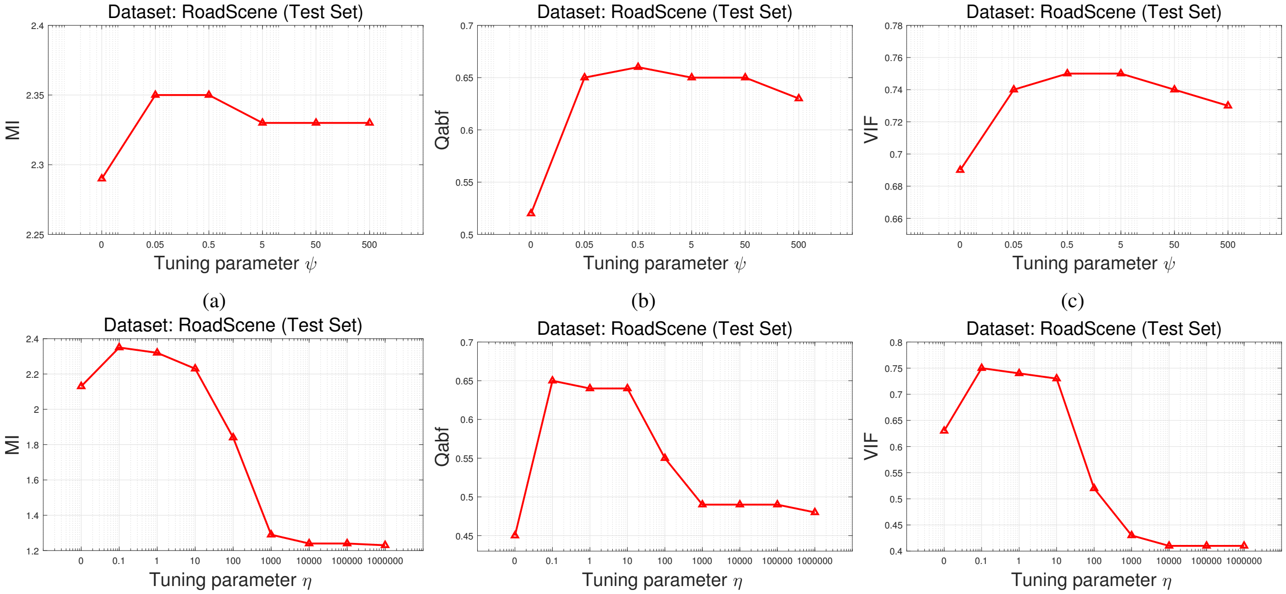
[1] https://github.com/Linfeng-Tang/MSRS

[2] https://github.com/hanna-xu/RoadScene

[3] https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029

[4] https://github.com/JinyuanLiu-CV/TarDAL

[5] http://www.med.harvard.edu/AANLIB/home.html



(a) (b) (c)

(d) (e) (f)

Figure 1: Fusion results among different configurations of DDFM on the test set. (a)-(c): $\eta = 0.1$ and $\psi = 0,\ 0.05,\ 0.5,\ 5,\ 50,\ 500$. (d)-(f): $\psi = 0.5$ and $\eta = 0,\ 0.1,\ 1,\ 10,\ 100,\ 1e3,\ 1e4,\ 1e5,\ 1e6$.

4. Selection for the hyperparameters

For our proposed DDFM, the tuning parameter ψ and η are important in the effectiveness of fusion. We show the results among different configurations via grid search on the RoadScene test set in Fig. 1. The metrics MI, Qabf and VIF are employed to determine the hyperparameters.

We first fix $\eta = 0.1$, and calculate the fusion quality when $\psi = 0,\ 0.05,\ 0.5,\ 5,\ 50,\ 500$ in Figs. 1a to 1c. Then we verify the fusion results for $\eta = 0,\ 0.1,\ 1,\ 10,\ 100,\ 1e3,\ 1e4,\ 1e5,\ 1e6$ when fixing $\psi = 0.5$ in Figs. 1d to 1f.

The change of ψ does not affect the results much, but the results are relatively high at $\psi = 0.5$. For the selection of η, it is clear that the performance of DDFM is the best when η is set to 0.1. A lower η causes a lack of detailed information, and a larger η makes the generation prior a higher place in fusion, generating artifacts unrelated to the original images.

Finally, to have a good fusion result, we set $\{\psi = 0.5, \eta = 0.1\}$ for the other experiments.

5. More qualitative comparison fusion results

More qualitative comparisons for *Infrared-Visible image Fusion* results are displayed in Figs. 2 to 4. Our method better integrates thermal radiation information in infrared images and detailed textures in visible images. Objects in dark regions are clearly highlighted, so that foreground targets can be easily distinguished from the background. Additionally, background details that are difficult to identify due to the low illumination have clear edges and abundant contour information, which help us understand the scene better.

More qualitative comparisons for *Medical Image Fusion* results are shown in Fig. 5. DDFM can better preserve the detailed texture and highlight the structure information than other methods.

References

[1] Harvard Medical website. http://www.med.harvard.edu/AANLIB/home.html. 3

[2] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *CVPR*, pages 5792–5801. IEEE, 2022. 3

[3] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1

[4] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion*, 83-84:79–92, 2022. 3

[5] Alexander Toet and Maarten A. Hogervorst. Progress in color night vision. *Optical Engineering*, 51(1):1 – 20, 2012. 3

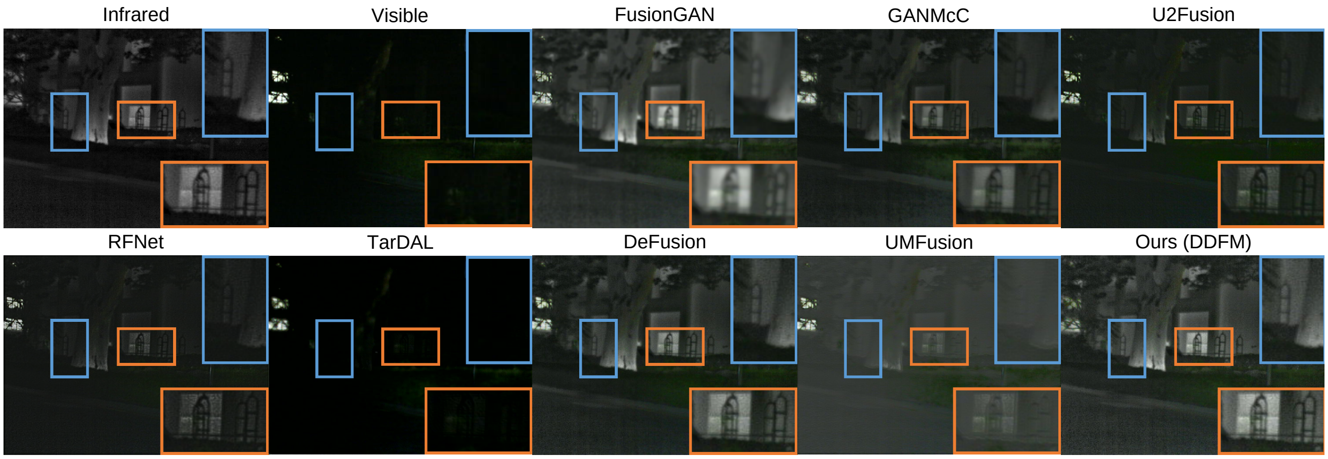


Figure 2: Qualitative results for *Infrared-Visible image Fusion* on MSRS dataset.

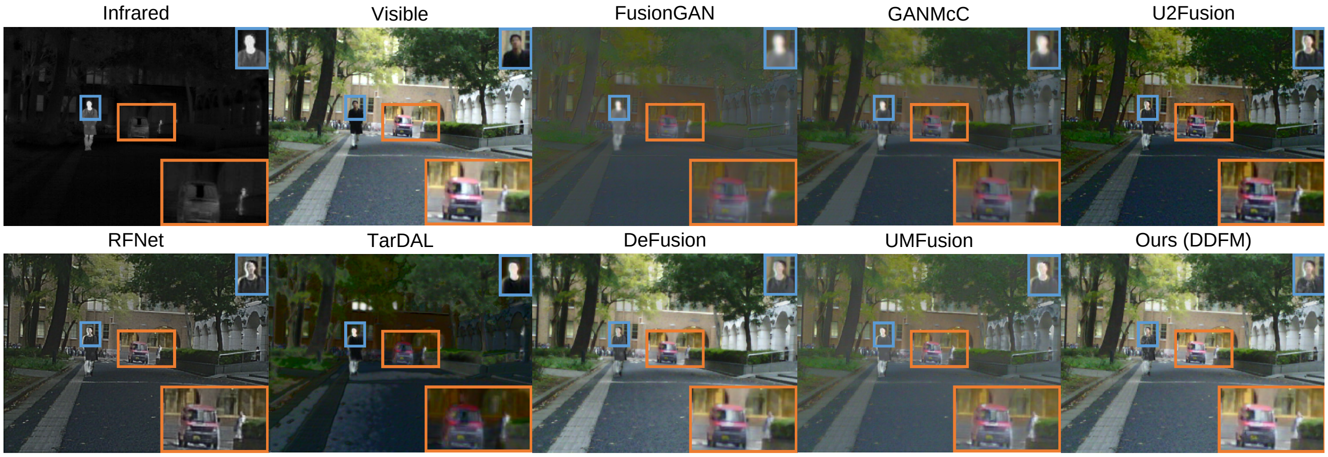


Figure 3: Qualitative results for *Infrared-Visible image Fusion* on MSRS dataset.

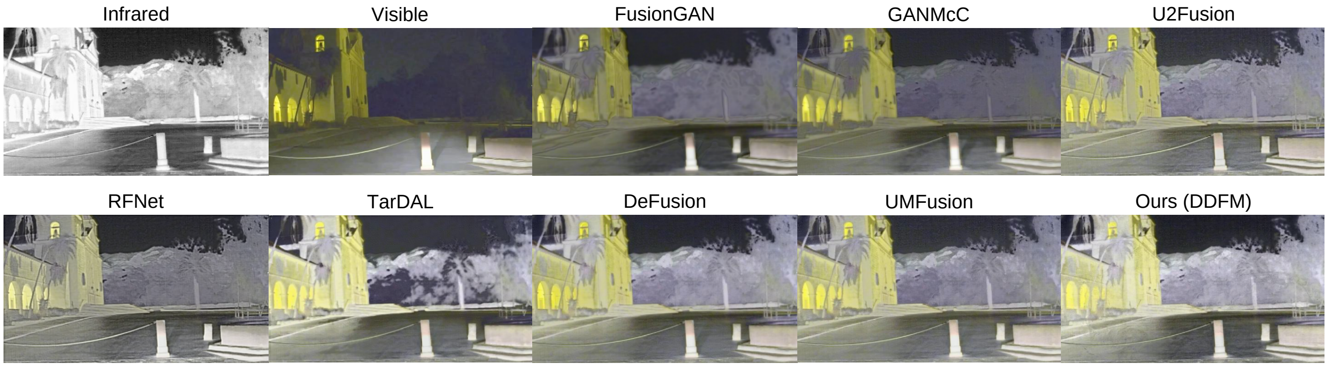


Figure 4: Qualitative results for *Infrared-Visible image Fusion* on RoadScene dataset.

[6] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *AAAI Conference on Artificial Intelligence, AAAI*, pages 12484–12491, 2020. 3

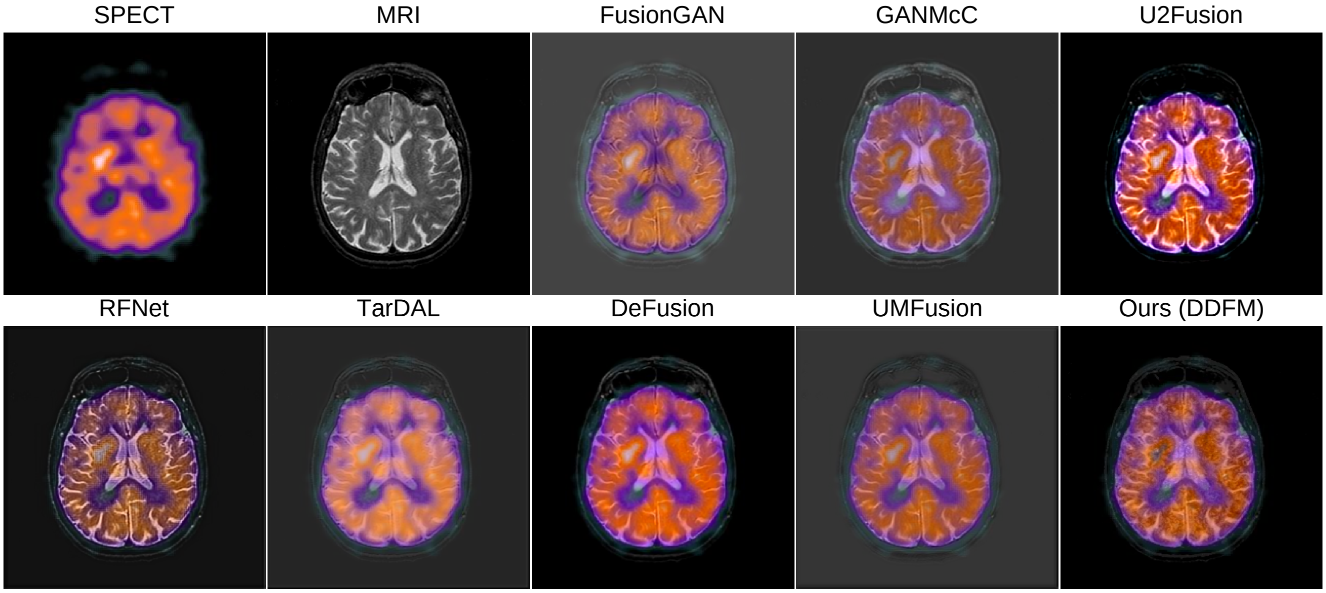


Figure 5: Qualitative results for *Medical Image Fusion* on Harvard Medical Image dataset.