Supplementary Material of TemCoCo

1. Comparison of Metrics

We propose two metrics flowD and feaCD in the main text to evaluate the temporal consistency of fused videos, from image-level and feature-level perspectives, respectively. The closest existing metrics to our proposed ones are tOF and LPIPS, which are widely used in the field of video super-resolution. Below, we provide the definitions of these two metrics. For the metric tOF:

$$E_{\mathrm{OF}}(t) = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{F}_{t\to t+1}(i) - \mathbf{F}_{\mathrm{gt},t\to t+1}(i)\|_2,$$
$$\mathrm{tOF} = \frac{1}{T-1}\sum_{t=1}^{T-1} E_{\mathrm{OF}}(t), \tag{1}$$

where $\mathbf{F}$ is the optical flow field, N denotes the number of pixels, and T represents the number of frames. For the metric LPIPS:

$$d_l(x,y) = \frac{1}{H_l W_l C_l}\sum_{h,w,c} \|\mathbf{f}_l(x)_{h,w,c} - \mathbf{f}_l(y)_{h,w,c}\|_2^2,$$
$$\mathrm{LPIPS}(x,y) = \sum_l w_l \cdot d_l(x,y),$$
$$\mathrm{LPIPS}_{\mathrm{video}} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{LPIPS}(x_t, y_t), \tag{2}$$

where $f_l(x)$ represents the feature of x at the l-th layer, and w_l denotes the corresponding weight.

The tOF metric assesses the temporal consistency of generated videos by computing the optical flow between the original and generated videos and then measuring their similarity. This is feasible in the video super-resolution field, as both the original and generated videos reside in the visible spectrum. However, in the video fusion task, no optical flow estimation method works universally across visible, infrared, and fused modalities. For example, using the Sea-RAFT model mentioned in our paper, Fig. 1 presents the estimated optical flow across different modalities, where the red box highlights regions with significant motion. Clearly, only the optical flow estimated from the visible modality accurately reflects the visual changes, which indicates that the tOF metric is not applicable to video fusion. Our proposed flowD adopts an indirect approach, predicting the optical flow solely on the visible video and using it to predict the next frame in the fused modality, thus avoiding optical flow estimation in the fused modality.

Regarding the LPIPS metric, it essentially measures the distance between the fused image and the original image in the feature space, without considering the temporal relationships in videos. In contrast, our proposed feaCD improves upon this approach. Instead of focusing solely on the similarity of images at the feature level, it emphasizes the consistency of directional changes in feature space across video frames.

2. Extension on Medical Video Fusion

We directly apply the proposed TemCoCo and comparative methods to medical video fusion on the *Harvard* dataset. The MRI and PET sequences in the *Harvard* dataset consist of a series of consecutive slice images, obtained from cross-sectional scans of the same organ. Due to the high correlation of anatomical structures between adjacent slices, which is analogous to the content changes between consecutive frames in a video, we treat each MRI or PET slice as a frame in a video for video fusion experiments. Figure 2 (a) presents a group of experimental results. The results demonstrate that only our method effectively preserves the structural and textural information of the MRI modality, while other methods exhibit unsatisfactory fusion outcomes. For instance, LRRNet, DDFM, and RCVS show poor contrast, DATFuse displays noticeable artifacts, and other methods also fail to adequately preserve texture and structure. Figure 2 (b) presents a set of temporal consistency evaluation results of medical video fusion. As highlighted in the red box, our method demonstrates the highest level of smoothness, whereas other methods exhibit varying degrees of artifacts. This verifies that our method also achieves the highest temporal consistency in medical video fusion.

3. Video Results Demonstration

We include additional dynamic video results in the supplementary material, covering both video fusion and video segmentation. For video fusion results, we recommend focusing more on the background regions to better observe the flickering phenomenon. For video segmentation results,

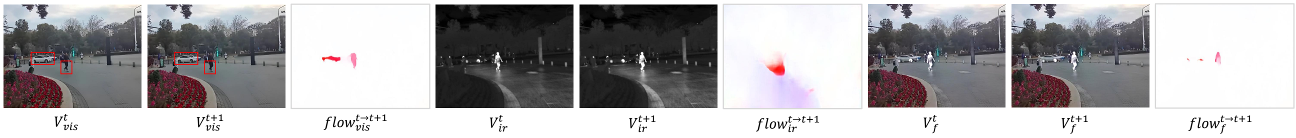
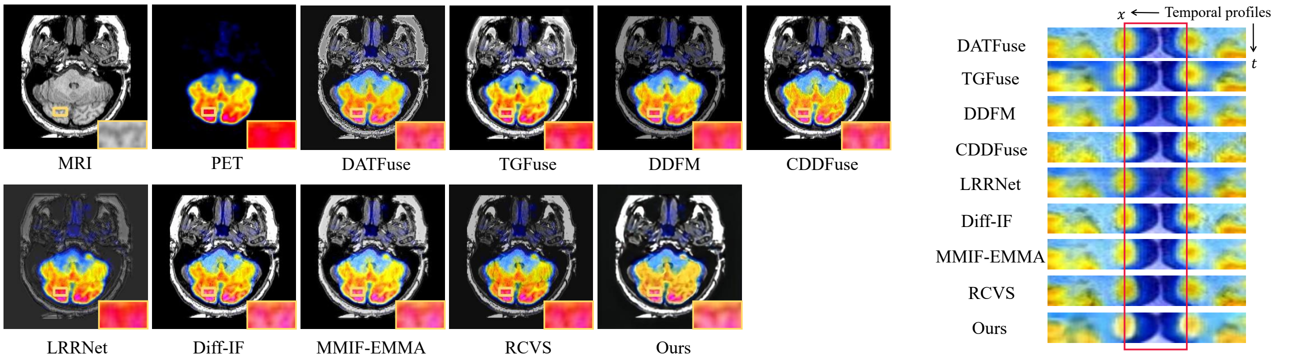


Figure 1. Flow calculation on different modalities.



(a) An example of medical video fusion results.

(b) Temporal results of medical video fusion.

Figure 2. An example of medical video fusion results.

we suggest paying attention to the accuracy of target segmentation, such as instances of mis-segmentation or missed segmentation. It is noteworthy that the video fusion results on the M3SVD dataset exhibit the most pronounced flickering artifacts with LRRNet and DDFM. These two methods also perform the worst on our proposed metrics, flowD and feaCD, which further validates that our proposed metrics are consistent with visual perception.