

# High Spatial-Temporal Resolution MRI Reconstruction for Speech Production Analysis

Qingyue Wei  
Department of Electrical Engineering  
Stanford University  
qywei@stanford.edu

Xuetong Zhou\*  
Department of Bioengineering  
Stanford University  
xuetongz@stanford.edu

## Abstract

*Real-time magnetic resonance imaging is widely used for analyzing structures during speech production. However, the trade-off between high spatial resolution and high temporal resolution always exists in acquiring MRI videos. In this project, our goal is to reconstruct high spatial-temporal resolution MRI videos based on the undersampled videos in both spatial and time domain. And a public multi-speaker dataset of speech production RT-MRI is used in this project [10]. To achieve the goal, we propose a pipeline that first applies a spatial reconstruction model with channel-wise attention mechanism to restore the high spatial resolution for each frame and then synthesize the missing frames by another optical flow-based framework using coarse to fine strategy. Experiments show that our spatial reconstruction model could outperform ADMM+TV [1] by a large margin under different undersampled ratios. And our proposed method also shows promising results for the reconstructed high spatial-temporal videos (40.87dB in PSNR). Moreover, the proposed method is also robust to different spatial undersampled input videos.*

## 1. Introduction

Human speech production involves a series of well-coordinated choreography of various soft-tissue and muscular structures, as well as bones [11]. Study of speech production is not only worthwhile for linguistic and speech science but can potentially facilitate the planning and evaluation of clinical treatments and surgeries [12]. Real-time magnetic resonance imaging (RT-MRI) is one of the modalities used to image the structures involved in speech production with the advantages of noninvasiveness and the ability to image arbitrary planes and deep structures [11] [12].

To achieve a high temporal resolution so that full movements of structures during the speech tasks can be captured, non-Cartesian fast imaging trajectories, such as radial and

spiral trajectories, and high undersampling rates are usually used to reduce the acquisition time of each 2D frame [16]. However, as a trade-off, spatial resolution is sacrificed when sampling with a higher temporal resolution [12]. In addition, the images would be likely to contain residual artifacts that are caused by a high undersampling rate. Based on these problems in speech production RT-MRI, the aim of this project is to develop a reconstruction method that provides speech production images with high spatial and temporal resolution through deep convolutional neural networks.

In this project, our goal is to reconstruct the MR images for each spatially undersampled frame and interpolate the missing frames to restore temporal resolution as well, given both spatial and temporal downsampled MRI k-space data. Inputs are RT-MRI frames that are undersampled in both spatial and temporal domains. We first use MICCAN to reconstruct the spatially undersampled images, then use IFNet to interpolate the missing frames. The outputs of our algorithm is the reconstructed high spatial and temporal resolution frames.

## 2. Related Work

Considering the particular acquisition method for MRI, reconstructing high spatial-temporal resolution MRI videos are treated as two different tasks in our proposed method. And the related work is also represented from two different aspects.

### 2.1. MRI spatial reconstruction

Recently, deep learning (DL) based methods are widely used for solving compressed sensing problems and show very promising reconstructed results. Popular DL-based methods could be divided into two groups *i.e.*, the end-to-end method, and unrolling method. For the first group, related methods mainly focus on applying various convolutional neural networks to learn the reconstructed outputs in an end-to-end way. Lee *et al.* [16] develop a multi-scale

deep neural network to reduce the aliasing artifacts. Han *et al.* [5] employ the domain adaption that could finetune on a model pre-trained on computed tomography (CT) images to restore the high spatial resolution MR images. Deep-complexmri [24] applies a complex convolutional network that could extract information from the real and imaginary parts of MR images simultaneously and their correlated relationships as well. As for the unrolling group, methods are used to unroll the network pipeline and use stages to simulate iterative optimizations. For instance, ADMM-Net [23] combines the Alternating Direction Method of Multipliers (ADMM) algorithm with a deep-learning model for MR image reconstruction problems. And Hammernik *et al.* [4] employ a DL-based variational network for MR image restoration. But most of these methods in the two groups only focus on 2D images. Relationships among 3D MR image slices are ignored. The unrolling methods using a specific iterative algorithm have higher interpretability than end-to-end models and end-to-end methods usually require more training data and higher computational power [9]. But on the other hand, end-to-end methods don't need any domain knowledge of MRI [9]. And recently, there are also end-to-end models using cascaded blocks to simulate the iterative reconstruction [21, 22] and show promising performances.

## 2.2. Video interpolation

Video interpolation is always a challenging field due to the non-linear motions and illumination changes in the videos [8]. Current widely used methods could be divided into two groups *i.e.*, the flow-based method and flow-free method. The flow-based method targets to estimate the intermediate flow that indicates the motion of each pixel. For example, Liu *et al.* [13] propose an end-to-end deep voxel flow (DVF) framework that could copy pixels nearby to get a more accurate flow estimation. And Park *et al.* [20] develop a model that applies bilateral motion for video interpolation which could directly estimate intermediate motions for video interpolation. In [25], quadratic interpolation is used to pay attention to the acceleration information shown in the consecutive frames and produce more accurate interpolation. However, the flow-based methods often require two steps *i.e.*, optical flow estimation and then intermediate frame synthesis. They rely heavily on the accuracy of the flow estimations and often require higher computation cost. The flow-free methods don't depend on optical flow estimation. For instance, Niklaus *et al.* [18] develop a fully convolutional network that could use several pairs of 1-D kernels for separable local convolution and directly generate the synthesized frame. And Phasenet [15] uses its own decoder to directly estimate the phase composition of the intermediate frame. In [2], PixelShuffle with channel attention is proposed to replace the flow estimation module.

## 3. Methods

### 3.1. Overview

Here, we propose a pipeline that could produce reconstructed high spatial-temporal resolution MRI videos based on the low spatial-temporal resolution data. As shown in Fig. 1, our pipeline is composed of two successive stages *i.e.*, spatial reconstruction and temporal interpolation. To be specific, given a pair of consecutive spatial and temporal undersampled MRI frames  $I_0, I_1$ , and a target timestep  $t$  where  $I_0, I_1 \in \mathbb{R}^{C \times H \times W}$ ,  $0 < t < 1$ , we will first restore corresponding high spatial resolution MRI images  $I_0^{rec}, I_1^{rec}$  and then use these two restored frames to synthesize the intermediate frame  $\tilde{I}_t$ . After applying such a pipeline through the whole MRI video, we can finally reconstruct a high spatial-temporal resolution video.

### 3.2. Spatial Reconstruction

To reconstruct MR images from the undersampled inputs, we follow the idea of [7] and adopt the MRI Cascaded Channel-wise Attention Network (MICCAN) for spatial reconstruction. MICCAN is an end-to-end network and uses cascaded blocks to simulate the iterative process in the unrolling method (*e.g.*, ADMM-Net [23]). As shown in Fig. 1(a), each cascaded block is composed of two modules *i.e.*, UNet-based Channel-wise Attention (UCA) module and Data Consistency (DC) module. The whole spatial reconstruction process could be represented as

$$I_i^{rec} = F_K^S(F_{K-1}^S(\dots(F_1^S(I_i, m_i; \theta_1^S)\dots); \theta_{K-1}^S); \theta_K^S), \quad (1)$$

where  $I_i$  is the  $i^{th}$  undersampled frame,  $I_i^{rec}$  is corresponding spatial reconstructed image and  $m_i$  is the undersampling mask for  $I_i$ .  $F_k^S(\cdot)$  is the  $k^{th}$  block and  $\theta_k^S$  is the corresponding parameters where  $k = 1, 2, \dots, K$ .  $K$  is the total number of the cascaded blocks. And below are more detailed explanations of each cascaded block.

**UNet-based Channel-wise Attention.** UCA is a convolution based module and follows an encoder-decoder structure combined with the channel-wise attention mechanism in the decoder part. The detailed structure of UCA is demonstrated in the appendix in Fig. 8. Considering that different channels of the feature maps contain diverse information, some channels could be important for reconstruction while others hardly contribute to it. It would be more reasonable if we treat them differently. Thus, a channel-wise attention mechanism is adopted to evaluate the importance of each channel and produce the channel-weighted feature maps. As shown in Fig. 2, similar to [6], the global average pooling is first applied to reduce the feature map size and

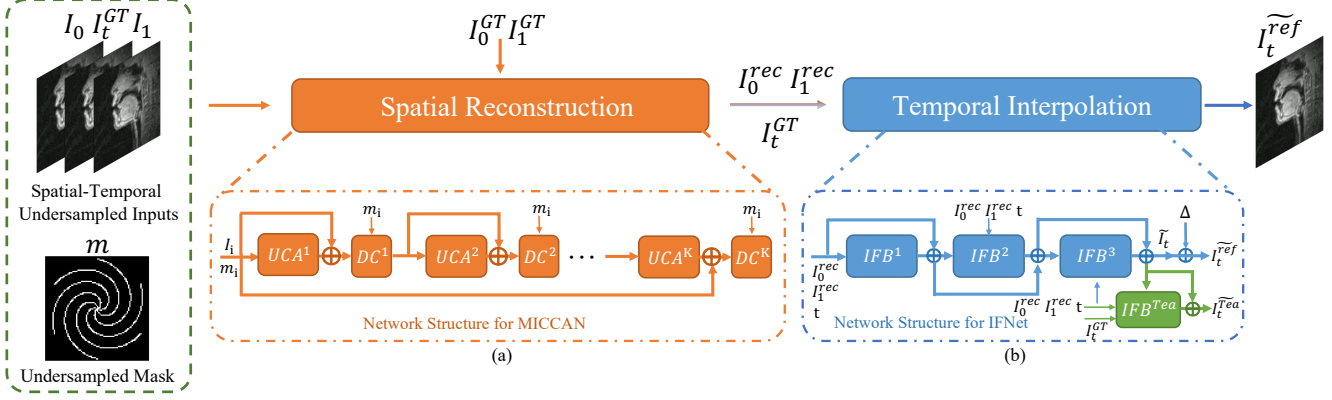


Figure 1. High spatial-temporal resolution reconstruction pipeline overview.

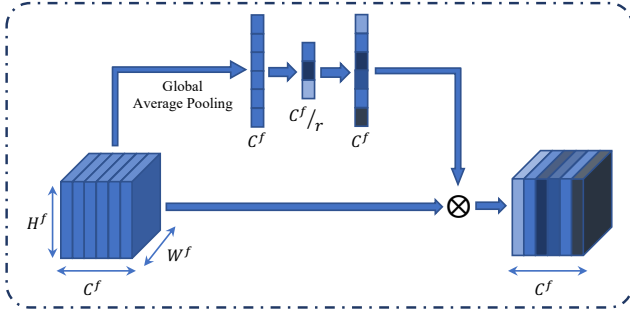


Figure 2. Channel-wise attention mechanism.

extract global information for each channel that

$$u_c = \frac{1}{H^f \times W^f} \sum_{p=1}^{H^f} \sum_{q=1}^{W^f} fea_{c,p,q}, \quad (2)$$

where  $u \in \mathbb{R}^{C^f}$  and  $c = 1, 2, \dots, C^f$  and  $fea \in \mathbb{R}^{C^f \times H^f \times W^f}$  is the input feature maps. Then, unlike SE block [6], here we use  $1 \times 1$  convolutional layers instead of fully connected layers to produce the channel weights for  $fea$  that

$$w = \sigma(G_2(\text{ReLU}(G_1(u)))), \quad (3)$$

where  $w \in \mathbb{R}^{C^f}$  is the channel weights.  $G_1(\cdot), G_2(\cdot)$  are both  $1 \times 1$  convolutional layers.  $G_1(\cdot)$  uses  $C^f, C^f/r$  as the input and output channel size respectively while  $G_2(\cdot)$  uses  $C^f/r, C^f$  as the input and output channel size respectively.  $\text{ReLU}(\cdot)$  indicates the ReLU activation function [3]. And then, the weighted feature maps could be written as

$$fea_w = w \circ fea, \quad (4)$$

where  $fea_w \in \mathbb{R}^{C^f \times H^f \times W^f}$  and  $\circ$  indicates the Hadamard product.

**Data Consistency.** To prevent the loss of information while the network goes deeper, we also use a skip connection before the DC module. Then, the DC module is employed to avoid the information degeneration and enforce the data fidelity between the reconstructed image  $I_i^{rec}$  and the undersampled frame  $I_i$  as well. This strategy could make sure that the sampled pixel positions in the reconstruction process have the same value as the undersampled input in k-space. Specifically,

$$\tilde{I}_k^{rec} = I_{k-1}^{rec} + I_k^{UCA}, \quad (5)$$

$$I_k^{rec} = f^{-1}(f(\tilde{I}_k^{rec}) + (f(I_i) - f(\tilde{I}_k^{rec})) \circ m_i), \quad (6)$$

where  $I_k^{UCA}$  is the output of the  $k^{th}$  UCA module and  $I_{k-1}^{rec}$  is the output of the  $(k-1)^{th}$  cascaded block.  $f$  indicates the fast Fourier transform,  $f^{-1}$  is the inverse fast Fourier transform and  $I_i$  is the undersampled input,  $m_i$  is the undersampling mask for  $I_i$  and  $\circ$  means the Hadamard product.

**Long-term skip connection.** Notice that in Fig. 1(a), the last cascaded block is slightly different from previous ones where we add another long-term skip connection between  $I_K^{UCA}$  and the input data  $I_i$ . This long skip connection aims to avoid the degeneration of the low frequency information and add more robustness to the network [7]. To be specific,

$$\tilde{I}_K^{rec} = I_i + I_K^{UCA}, \quad (7)$$

where  $\tilde{I}_K^{rec}$  is the input of the last DC module.

**Overall Loss.** To preserve the high frequency information, a combined loss is used for spatial reconstruction that

$$\begin{aligned} \mathcal{L}_{spa}^{rec} &= \mathcal{L}_1 + \lambda_p \mathcal{L}_{per}, \\ \mathcal{L}_1 &= \|I_i^{GT} - I_i^{rec}\|_1, \\ \mathcal{L}_{per} &= \sum_{l=1}^L \|\phi_{VGG}^l(I_i^{GT}) - \phi_{VGG}^l(I_i^{rec})\|_2^2, \end{aligned} \quad (8)$$

where  $I_i^{GT}$  is the ground-truth MR image of  $I_i$ ,  $\mathcal{L}_1$  indicates the  $l_1$  loss and  $\mathcal{L}_{per}$  is the perceptual loss using  $l_2$  distance and a pretrained VGG16 model. It could force the intermediate feature maps at different scales of  $I_i^{GT}$ ,  $I_i^{ec}$  from the pretrained model to be close and thus their contents and global structures are forced to be similar and  $\lambda_p$  is the weight for  $\mathcal{L}_{per}$ .

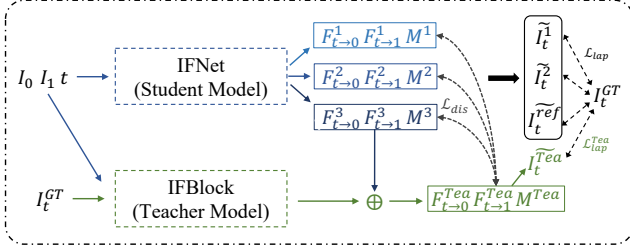


Figure 3. IFNet pipeline overview.

### 3.3. Temporal Interpolation

As for temporal interpolation, we follow the idea of IFNet [8] which could directly synthesize the intermediate frame from coarse to fine with less time cost. The whole pipeline of temporal interpolation is shown in Fig. 1(b) and we also demonstrate a more detailed pipeline in Fig. 3. Specifically, given two consecutive frames  $I_0, I_1 \in \mathbb{R}^{C \times H \times W}$  and the target timestep  $t$ ,  $t$  will first be encoded as a separate input channel  $t^{TE}$  where  $t^{TE}$  has the same shape of  $I_0$  and  $t_{i,j}^{TE} = t$  for every position in  $t^{TE}$ .  $I_0, I_1$  together with  $t^{TE}$  will be fed into IFNet to get the estimated intermediate optical flow  $F_{t \to 0}, F_{t \to 1}$  which could be used to estimate the pixel-level motion. Besides the flow, a fusion map  $M$  is estimated for a better combination of  $F_{t \to 0}, F_{t \to 1}$ . And these whole estimation process follows a coarse-to-fine strategy that it will start from the low spatial resolution to make coarse estimation and then switch to higher spatial resolution for finer estimation. Under this pattern, it could first detect larger motions and then gradually focus more on tiny movements. This coarse-to-fine process is simulated by three cascaded IFBlocks which are composed of several convolutional layers with specific downsampled scales and a transposed convolutional layer. More detailed structure could be found in the appendix in Fig. 9. Concretely, for each IFBlock, we have

$$F_{t \to 0}^m, F_{t \to 1}^m, M^m = IFB^m(I_0, I_1, \hat{I}_{t \to 0}^{m-1}, \hat{I}_{t \to 1}^{m-1}, t^{TE}, F_{t \to 0}^{sum_{m-1}}, F_{t \to 1}^{sum_{m-1}}, M^{sum_{m-1}}), \quad (9)$$

where

$$\hat{I}_{t \to 0}^{m-1} = \mathcal{W}(I_0, F_{t \to 0}^{m-1}), \hat{I}_{t \to 1}^{m-1} = \mathcal{W}(I_1, F_{t \to 1}^{m-1}), \quad (10)$$

$$\begin{aligned} F_{t \to 0}^{sum_m} &= F_{t \to 0}^{sum_{m-1}} + F_{t \to 0}^m, \\ F_{t \to 1}^{sum_m} &= F_{t \to 1}^{sum_{m-1}} + F_{t \to 1}^m, \\ M^{sum_m} &= M^{sum_{m-1}} + M^m, \end{aligned} \quad (11)$$

$IFB^m$  represents the  $m^{th}$  IFBlock and  $m = 1, 2, 3$ .  $F_{t \to 0}^m, F_{t \to 1}^m, M^m$  indicate the estimated intermediate optical flows and the fusion map from the  $m^{th}$  IFBlock.  $\mathcal{W}$  denotes the image backward warping process.  $F_{t \to 0}^{sum_m}, F_{t \to 1}^{sum_m}, M^{sum_m}$  are the accumulated optical flows and the fusion map. Notice that, for the first IFBlock, the inputs are  $I_0, I_1, t^{TE}$  only. And all inputs are concatenated along the channel dimension and then feed into the IFBlock. Then, the estimated intermediate frame  $\tilde{I}_t^m$  is calculated as

$$\tilde{I}_t^m = M^m \circ \hat{I}_{t \to 0}^m + (1 - M^m) \circ \hat{I}_{t \to 1}^m, \quad (12)$$

where  $\circ$  denotes the element-wise multiplication.

Moreover, as shown in Fig. 3, to get better intermediate estimations and have more supervision, a privileged distillation structure is applied during training process. Specifically, all inputs and the ground-truth frame  $I_t^{GT}$  for  $\tilde{I}_t$  will be fed into another IFBlock called Teacher Model denoted as  $IFB^{Tea}$  to get the estimated flow output and fusion map that

$$F_{t \to 0}^{Tea}, F_{t \to 1}^{Tea}, M^{Tea} = IFB^{Tea}(I_0, I_1, \hat{I}_{t \to 0}^3, \hat{I}_{t \to 1}^3, t^{TE}, F_{t \to 0}^{sum_3}, F_{t \to 1}^{sum_3}, M^{sum_3}, I_t^{GT}), \quad (13)$$

$F_{t \to 0}^{Tea}, F_{t \to 1}^{Tea}, M^{Tea}$  will then be added with the corresponding outputs from the last IFBlock of the student model separately to get  $\hat{I}_{t \to 0}^{Tea}, \hat{I}_{t \to 1}^{Tea}, M^{sum}$  and  $\tilde{I}_t^{Tea}$ . Then  $F_{t \to 0}^{Tea}, F_{t \to 1}^{Tea}$  will be used to get a distillation loss where

$$\mathcal{L}_{dis} = \sum_{m=1}^3 \sum_{i=0,1} \|F_{t \to i}^{Tea} - F_{t \to i}^m\|_1. \quad (14)$$

Moreover, a UNet-based encoder-decoder structure is employed to reduce the artifacts and rectify the high-frequency information of  $\tilde{I}_t^3$ . To be specific, first, we use another cascaded convolutional layers to extract contextual features in a pyramid way where the inputs are  $I_0/I_1$  and the corresponding accumulated flow from the student model.  $I_0/I_1$  and the flow would be downsampled by different scales and then use backward warping to get estimated intermediate frames at different scales. And then together with  $I_0, I_1, \hat{I}_{t \to 0}^3, \hat{I}_{t \to 1}^3, F_{t \to 0}^{sum_3}, F_{t \to 1}^{sum_3}, M^{sum_3}$  would be fed into the UNet-based encoder-decoder structure to get an estimated correction term, denoted as  $\Delta$ . Then we can get the refined output as

$$\tilde{I}_t^{ref} = \tilde{I}_t^3 + \Delta. \quad (15)$$

The whole loss for the frame interpolation training process

could then be described as

$$\begin{aligned}\mathcal{L}_{temp}^{int} &= \mathcal{L}_{lap} + \mathcal{L}_{lap}^{Tea} + \lambda_d \mathcal{L}_{dis}, \\ \mathcal{L}_{lap} &= lap(\tilde{I}_t^{ref}, I_t^{GT}) + \sum_{m=1}^2 lap(\tilde{I}_t^m, I_t^{GT}), \\ \mathcal{L}_{lap}^{Tea} &= lap(\tilde{I}_t^{Tea}, I_t^{GT}),\end{aligned}\quad (16)$$

where  $lap(\cdot)$  means the laplacian loss for evaluating frame interpolation [17] and  $\lambda_d$  is the weight for the distillation loss.

Code for MICCAN and IFNet models are adapted from <https://github.com/charwing10/isbi2019miccan> and <https://github.com/hzwer/arXiv2021-RIFE>. And training process also follows their patterns but has been made a lot of changes to suit our own inputs, undersampled patterns, evaluations, and so on. Moreover, code related to the data process and combination of these two stages are all written by ourselves.

## 4. Dataset

A multi-speaker dataset of speech production RT-MRI is used in this project [10]. The dataset includes raw data and synchronized audio acquired during multiple speaking tasks on 75 speakers. Each speaker includes 20-24 speaking videos and the duration of each video is around 20-40 seconds. Each video has a frame rate at 83 fps. We split all of the speakers into three datasets *i.e.*, training, validation and test set. 60 speakers are included in the training set while the other 15 speakers are shared by validation and test set. The total frames of the raw data is above 2 millions. Thus, considering the time consuming and the computation limitation, we only randomly take one video for each speaker in training set and two different videos of each speaker in the remained 15 speakers for validation and test set respectively. And we also 2X downsample the video along time dimension and only take the half time duration of each video from the middle. Each frame has a shape of  $84 \times 84$ . The total frame number of the training/validation/test set is around 40K/10K/10K. For both spatial and temporal reconstruction, all inputs are normalized between 0 and 1 range. Moreover, as for augmentation, we adopt horizontal flip, vertical flip,  $90^\circ$  clock-rotation,  $90^\circ$  counter clock-rotation. And these augmentations are randomly picked during training. Notice that, for temporal interpolation, two consecutive frames are applied with the same augmentation strategy. And three consecutive frames from one video in our dataset are shown in Fig. 4(a).

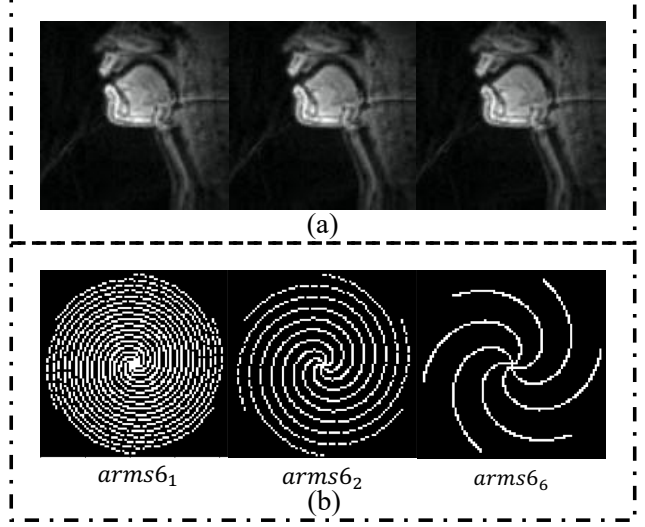


Figure 4. Sample frames of the speak video dataset and the spatial undersampling patterns.

## 5. Experiments and Discussion

### 5.1. Experiment details

For spatial reconstruction, we set  $K = 5$  so that we have 5 cascaded blocks in total.  $\lambda_p$  for the perceptual loss is set to be 0.1. And we apply Adam to optimize network parameters with the learning rate at  $1e^{-4}$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  and  $\epsilon = 1e^{-8}$ . And the decay rate is at 0.5 with milestone epochs at 5, 20, 35 separately. We train the spatial reconstruction network for 100 epochs. As for temporal interpolation, we choose  $[4, 2, 1]$  as the downsampled scales for  $IFB^1, IFB^2, IFB^3$  separately and for the teacher model, the downsampled scale is set to be 1.  $\lambda_d$  for the distillation loss is set as 0.01. And we use  $t = 0.5$  for all interpolation experiments in this project. We apply cosine annealing strategy to gradually decrease the learning rate from  $3e^{-4}$  to  $3e^{-5}$ . And AdamW [14] is used to optimize parameters of IFNet. And we train the temporal interpolation network for 200 epochs. Moreover, these two sets of learning rates mentioned above are recommend by their original papers and also tested on our own input data for a few epochs and showed reasonable decreasing loss. Inputs for both spatial and temporal reconstruction are resized to  $128 \times 128$ . To take full advantage of the GPU memory, we set the batch size for spatial reconstruction as 32 and 144 for temporal interpolation. And we pick the best models for the spatial and temporal reconstruction based on their performances on the validation set and then test them on the test set after training. All experiments are implemented in PyTorch. As for evaluation metrics, we adopt Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) to evaluate the performances of the proposed method.

## 5.2. Undersampled MRI Synthesis

The original RT-MRI images included by the dataset were acquired by a 13-interleaf spiral-out spoiled gradient-echo pulse sequence [11]. In each repetition time, one readout interleaf starts from the center of k-space with a different initial angle and spirals out. The 13 interleaves are collected together as one k-space coverage to reconstruct one frame of image. Interleaved spiral pulse sequence is a widely used fast imaging sequence in MR, especially in fields such as RT-MRI, MR fluoroscopy and cine imaging, for it could increase the temporal resolution and reduce motion artifacts. Therefore, in order to simulate the real-world MRI acquisition, we designed our undersampling pattern based on the interleaved spiral trajectory by reducing the number of arms. The sampling masks were generated by MRI function of SigPy [19]. We used 6 spiral interleaves, which is more than a half reduced compared to the original sampling pattern. The matrix size were set to  $84 \times 84$ , which is the same with the matrix size of the images. We also tried undersampling along each readout trajectory with rates of 1, 2, and 6, denoted as  $arms6_1$ ,  $arms6_2$ ,  $arms6_6$ . And the corresponding undersampled masks are shown in Fig. 4(b). The undersampling was applied retrospectively by multiplication in k-space.

## 5.3. Spatial Reconstruction

**Ablation study on MICCAN.** To explore how different components contribute to the performance of the spatial reconstruction model MICCAN, we conduct the following experiments using the undersampled ratio  $arms6_2$ . 1) **MICCAN with only UNet.** This means that in the UNet-based channel-wise attention module, we don't include the channel-wise attention mechanism and only keep the UNet-based encode-decoder structure. Also, for the last cascade block, we don't apply long-term skip connection; 2) **MICCAN with only UCA**, which doesn't apply long-term skip connection for the last block; 3) **MICCAN with UCA and long-term** which adopt channel-wise attention in each block as well as a long skip connection at last. And this is also the final configuration of our spatial reconstruction model. The results are shown in Table 1. Notice that results for validation set and test set are shown simultaneously in all tables where the test set results are filled in the brackets. In Table 1, we could see that adding channel-wise attention mechanism doesn't show significant improvements and using long-term skip connection could only gain a very slight increase in PSNR and SSIM. We believe the reason for these might be that our original inputs don't have very good quality due to its low spatial-resolution and high undersampling rate. Therefore, our ground-truth has artifacts itself which might cause these components failing to contribute to the final results.

**Experiments on different undersampled ratios** To ex-

plore the robustness of our spatial reconstruction model under different undersampled ratios, we also conduct experiments based on the undersampled mask  $arms6_1$ ,  $arms6_2$ ,  $arms6_6$  respectively. Meanwhile, we also adopt the Alternating Direction Methods of Multipliers (ADMM) algorithm [1] combined with Total Variation (TV) as the regularizer as another baseline. ADMM is an iterative approach for solving convex optimization problems. As shown in Table 2, we could see that our model outperforms ADMM+TV by a very large margin *e.g.*, using undersampled mask  $arms6_1$ , our method yields another 9.7dB improvements in PSNR comparing with ADMM+TV. Moreover, when the undersampled ratio along each readout arm increase to 2, the performance only drops from 34.07dB to 29.64dB in PSNR which indicates our model is robust to deal with variations in undersampled patterns and high undersampling ratios as well.

**Qualitative Analysis.** We also demonstrate the reconstruction outputs of one specific input sample under different undersampled ratios (*i.e.*,  $arms6_1$ ,  $arms6_2$ ,  $arms6_6$ ) from our model as well as ADMM+TV with error maps shown together. In Fig. 5, we could see that while the undersampled ratio increase, performance of ADMM+TV drops rapidly, its reconstructed images for undersampling ratio  $arms6_6$  are very blur and lose lots structure and context details while our MICCAN could preserve a lot of detailed information (*e.g.*, edges, contrast and so on) even under a high undersampled ratio.

Table 1. Ablation study for spatial reconstruction on validation and test set.

| UNet | UCA | long-term | PSNR(dB)↑           | SSIM ↑                |
|------|-----|-----------|---------------------|-----------------------|
| ✓    |     |           | 29.57(29.58)        | 0.8314(0.8303)        |
|      | ✓   |           | 29.57(29.63)        | 0.8319(0.8321)        |
|      | ✓   | ✓         | <b>29.64(29.72)</b> | <b>0.8330(0.8340)</b> |

Table 2. Results for frame reconstruction on validation and test set under different undersampling ratios.

| Method  | Ratio    | PSNR(dB)↑    | SSIM ↑         |
|---------|----------|--------------|----------------|
| ADMM+TV | $arm6_1$ | 24.37(24.43) | 0.7830(0.7862) |
| Ours    | $arm6_1$ | 34.07(34.27) | 0.9157(0.9168) |
| ADMM+TV | $arm6_2$ | 20.32(20.30) | 0.6023(0.6032) |
| Ours    | $arm6_2$ | 29.64(29.72) | 0.8330(0.8340) |
| ADMM+TV | $arm6_6$ | 17.39(17.40) | 0.4910(0.4918) |
| Ours    | $arm6_6$ | 26.89(26.97) | 0.7630(0.7645) |

## 5.4. Temporal Interpolation

**Experiments on different undersampled spatial reconstruction.** For temporal interpolation, we conducts several experiments using the spatial reconstructed images under different undersampled ratios (*i.e.*,  $arms6_1$ ,  $arms6_2$ ,  $arms6_6$ ) to see the robustness of our

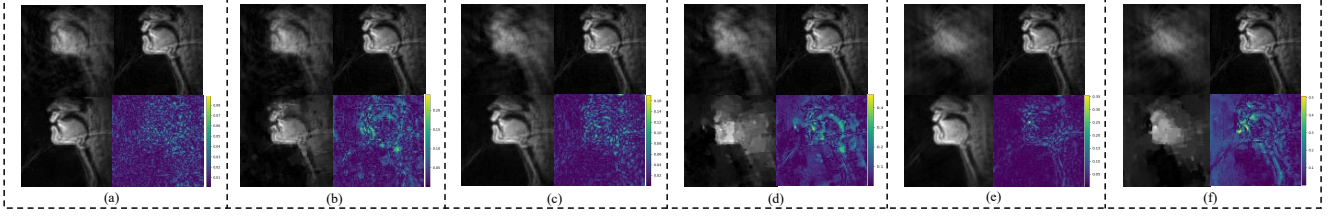


Figure 5. Spatial reconstruction visualization. For each dashed square, the undersampled input, ground-truth, reconstructed image, error map are shown from left-top to right-bottom respectively. Dashed square (a) shows results using MICCAN under  $arms6_1$ ; (b) shows results using ADMM+TV under  $arms6_1$ ; (c) shows results using MICCAN under  $arms6_2$ ; (d) shows results using ADMM+TV under  $arms6_2$ ; (e) shows results using MICCAN under  $arms6_6$ ; (f) shows results using ADMM+TV under  $arms6_6$ .

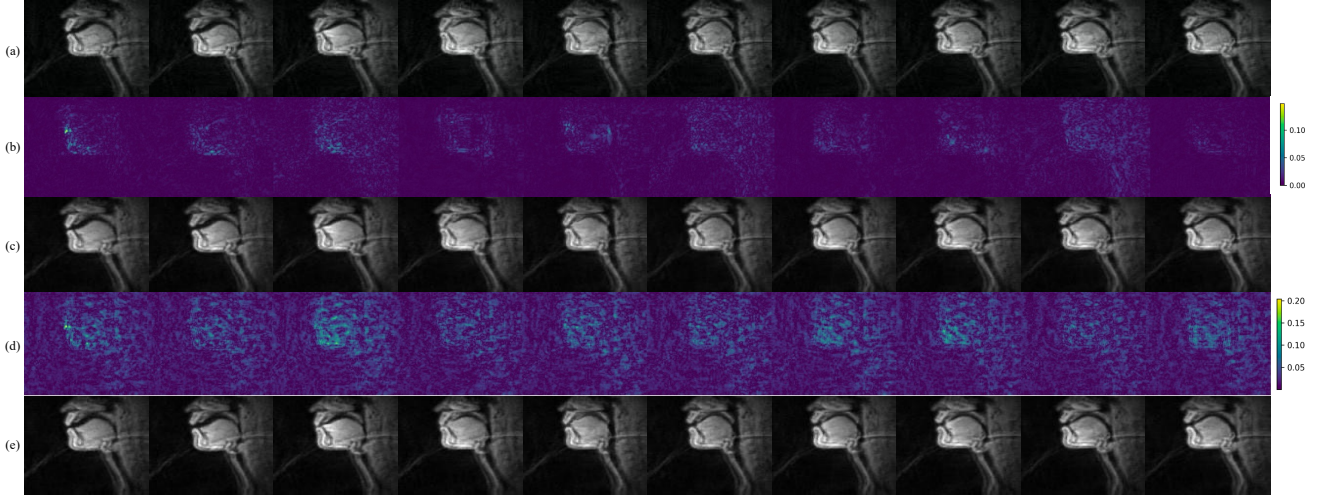


Figure 6. Temporal interpolation visualization. Here we show a fragment of a spatial reconstructed and temporal interpolated video. To enlarge the differences between two adjacent frames, the interval between two adjacent frames are set as 8 frames. And the spatial undersampled ratio is  $arms6_1$ . Row (a) shows the ground-truth frames that are directly spatially reconstructed from the corresponding spatial undersampled raw frames; Row (c) shows the corresponding synthesized frames based on the its front and back spatial reconstructed frames; Row (e) shows the real ground-truth frames from the raw data; Row (b) are the error maps between Row (a) and Row (c); Row (d) are the error maps between Row (e) and Row (c).

temporal interpolation model when inputs have different qualities. Moreover, we also include another experiment using the images from the raw data as a reference (denoted as original). As shown in Table 3,  $PSNR^{rec}$  indicates that it is calculated between the outputs and the reconstructed images (they are directly reconstructed by MICCAN using corresponding spatial undersampled raw data as inputs) while  $PSNR^{gt}$  means it's evaluated between the outputs and the original images from raw data. From Table 3, we could see that IFNet is quite robust to inputs with various qualities that  $PSNR^{rec}$  only varies from 40.87dB to 41.41dB under different undersampled spatial reconstruction inputs. One interesting finding is that under more severer undersampled ratio, the drop between the  $PSNR$  of inputs and  $PSNR^{gt}$  of outputs even decreases (3.85dB( $arms6_1$ ) to 2.82dB( $arms6_6$ ) on validation set).

**Qualitative Analysis.** We also show the interpolated out-

puts of one specific video that are first spatially reconstructed and then temporally interpolated. The spatial undersampled ratio is  $arms6_1$ . As shown in Fig. 6, we also demonstrate the reconstructed ground-truth in Row (a) as well as the real ground-truth in Row(e). From Row (b), we could notice that the error maps between the synthesized frames and its spatial reconstructed ground-truth are nearly close to zero. And there are only slight differences between the synthesized frames and the real ground-truth shown in Row (d) which indicates that IFNet could synthesize frames that could be considered as high-quality in the perspective of its training data. This means that it is the quality of the input data that constrains the output of IFNet. And we also demonstrated the tongue movements in one specific videos shown in Fig. 7. Here  $I_{gt}/I_{gt}^{rec}/I_{syn}$  indicate the tongue movements in the raw video/spatial reconstructed ground-truth video/spatial reconstructed and tem-

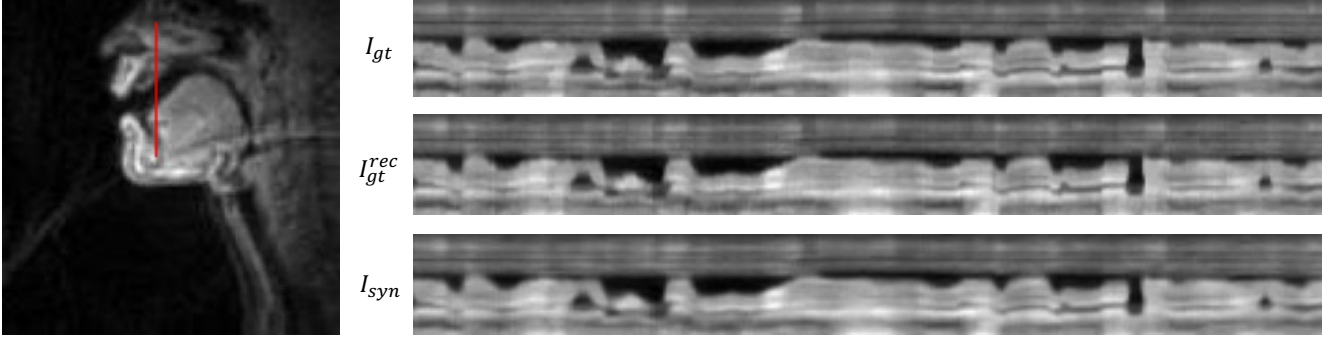


Figure 7. Tongue movements in one speaking video. Images are composed by the same column at different frames along the red line shown in the most left sample frame. And  $I_{gt}$  means the tongue movements in the raw video.  $I_{gt}^{rec}$  indicates tongue movements in the spatial reconstructed ground-truth video. And  $I_{syn}$  represents tongue movements in the temporal interpolated and spatial reconstructed video.

poral interpolated video respectively. Notice that tongue movements in  $I_{syn}$  is highly consistent with  $I_{gt}$  and have clear boundary. These mean that IFNet could correctly estimate the movements of the tongue between two consecutive frames. Moreover, increasing temporal undersampling ratio means that we could have more time to collect k-space data for each frame. Therefore, we could get much higher spatial resolution and also get larger image size. Then, we could do temporal interpolation based on frames in higher spatial resolution. And finally produce the synthesized videos with both higher spatial and temporal resolution.

For all conducted experiments, we think that our models haven't overfitted on the training set since we have monitored the training process and checked validation results (e.g., loss, PSNR). Validation loss didn't show a sign of overfitting. To prevent overfitting, we have adopted learning rate decay strategies for both models during training, used data augmentations, included dropout layers in the models and so on. Meanwhile, we could notice that all results of validation set and test set are very similar which might indicate that our method has a good generalization ability.

Table 3. Results for frame interpolation on validation and test set based on spatial reconstruction with different undersample ratios.

| Ratio              | PSNR <sup>rec</sup> (dB) ↑ | SSIM <sup>rec</sup> ↑ | PSNR <sup>gt</sup> (dB) ↑ | SSIM <sup>gt</sup> ↑ |
|--------------------|----------------------------|-----------------------|---------------------------|----------------------|
| arms6 <sub>1</sub> | 40.87(40.73)               | 0.9938(0.9936)        | 30.22(30.20)              | 0.9399(0.9410)       |
| arms6 <sub>2</sub> | 41.41(41.28)               | 0.9949(0.9947)        | 26.77(26.72)              | 0.8846(0.8854)       |
| arms6 <sub>6</sub> | 40.86(40.80)               | 0.9940(0.9938)        | 24.07(24.19)              | 0.8326(0.8348)       |
| original           | —                          | —                     | 39.33(39.22)              | 0.9928(0.9925)       |

## 6. Conclusion and Future Work

In this project, we propose a pipeline that could reconstruct high spatial-temporal resolution MRI videos. We split the task into two stages. Given a MRI video which are undersampled in both spatial and time domain, we first apply MICCAN to restore high spatial resolution for each frame and then adopt IFNet to synthesize the missing frames based

on the spatial reconstructed frames. And then we could produce a high spatial-temporal resolution MRI video. And we use a public multi-speaker dataset of speech production RT-MRI to evaluate our proposed method. For spatial reconstruction, our method could outperform the ADMM+TV by a large margin under different undersampled ratios. And the frame interpolation also achieve high PSNR and shows robustness to inputs with various qualities (i.e., spatially reconstructed videos under different spatial undersampling ratios). Qualitative analysis also show that we could accurately estimate the tongue movements in the speaking videos.

Considering that we only conduct experiments on one dataset, future work could include more datasets that have higher quality to investigate the performance and the robustness of our proposed method. For the spatial reconstruction, we could also come up a way to take the relationships among nearby frames into consideration. Moreover, in this project, we only synthesize one intermediate frame for  $t = 0.5$ . We could also increase interval between  $I_0, I_1$  which means higher undersampled ratio in time domain and try to do interpolation multiple frames for different  $t$ .

## 7. Appendices

More detailed structure for UCA module of MICCAN is shown in Fig. 8. And the detailed structure of IFBlock in IFNet is shown in Fig. 9

## 8. Contributions and Acknowledgements

**Code.** Code for MICCAN and IFNet models are adapted from <https://github.com/charwing10/isbi2019miccan> and <https://github.com/hzwer/arXiv2021-RIFE>.

**Contributions.** Q.W. conduct all the experiments & write the final report & make the poster for this project.

**Collaborator.** X.Z., who is a non-CS231N collabora-



- [12] Sajjan Goud Lingala, Yinghua Zhu, Yoon Chul Kim, Asterios Toutios, Shrikanth Narayanan, and Krishna S. Nayak. A fast and flexible MRI system for the study of dynamic vocal tract shaping. *Magnetic Resonance in Medicine*, 77(1):112–125, 2017. 1
- [13] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4463–4471, 2017. 2
- [14] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 5
- [15] Simone Meyer, Abdelaziz Djelouah, Brian McWilliams, Alexander Sorkine-Hornung, Markus Gross, and Christopher Schroers. Phasenet for video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 498–507, 2018. 2
- [16] Aaron Niebergall, Shuo Zhang, Esther Kunay, Götz Keydana, Michael Job, Martin Uecker, and Jens Frahm. Real-time mri of speaking at a resolution of 33 ms: undersampled radial flash with nonlinear inverse reconstruction. *Magnetic Resonance in Medicine*, 69(2):477–485, 2013. 1
- [17] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1710, 2018. 5
- [18] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 670–679, 2017. 2
- [19] Frank Ong and Michael Lustig. Sigpy: a python package for high performance iterative reconstruction. In *Proceedings of the ISMRM 27th Annual Meeting, Montreal, Quebec, Canada*, volume 4819, 2019. 6
- [20] Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation. In *European Conference on Computer Vision*, pages 109–125. Springer, 2020. 2
- [21] Chen Qin, Jo Schlemper, Jose Caballero, Anthony N Price, Joseph V Hajnal, and Daniel Rueckert. Convolutional recurrent neural networks for dynamic mr image reconstruction. *IEEE transactions on medical imaging*, 38(1):280–290, 2018. 2
- [22] Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for mr image reconstruction. In *International conference on information processing in medical imaging*, pages 647–658. Springer, 2017. 2
- [23] Jian Sun, Huibin Li, Zongben Xu, et al. Deep admm-net for compressive sensing mri. *Advances in neural information processing systems*, 29, 2016. 2
- [24] Shanshan Wang, Huitao Cheng, Leslie Ying, Taohui Xiao, Ziwen Ke, Hairong Zheng, and Dong Liang. Deepcomplexmri: Exploiting deep residual network for fast parallel mr imaging with complex convolution. *Magnetic Resonance Imaging*, 68:136–147, 2020. 2
- [25] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. *Advances in Neural Information Processing Systems*, 32, 2019. 2