

## Position Paper

## Neural collaborative filtering for unsupervised mitral valve segmentation in echocardiography

Luca Corinzia <sup>a,\*</sup>, Fabian Laumer <sup>a</sup>, Alessandro Candreva <sup>b</sup>, Maurizio Taramasso <sup>c</sup>, Francesco Maisano <sup>c</sup>, Joachim M. Buhmann <sup>a</sup>

<sup>a</sup> ETH Zurich, Institute for Machine Learning, Zurich, Switzerland

<sup>b</sup> University Hospital Zurich, Department of Cardiology, Zurich, Switzerland

<sup>c</sup> University Hospital Zurich, Department of Cardiac Surgery, Zurich, Switzerland

## ARTICLE INFO

**Keywords:**  
Mitral valve  
Segmentation  
Collaborative filtering  
Neural network

## ABSTRACT

The segmentation of the mitral valve annulus and leaflets specifies a crucial first step to establish a machine learning pipeline that can support physicians in performing multiple tasks, e.g. diagnosis of mitral valve diseases, surgical planning, and intraoperative procedures. Current methods for mitral valve segmentation on 2D echocardiography videos require extensive interaction with annotators and perform poorly on low-quality and noisy videos. We propose an automated and unsupervised method for the mitral valve segmentation based on a low dimensional embedding of the echocardiography videos using neural network collaborative filtering. The method is evaluated in a collection of echocardiography videos of patients with a variety of mitral valve diseases, and additionally on an independent test cohort. It outperforms state-of-the-art *unsupervised* and *supervised* methods on low-quality videos or in the case of sparse annotation.

## 1. Introduction

The mitral valve (MV) is the largest valve of the heart and safeguards the monodirectional blood flow from the left atrium towards the left ventricle. It is composed of two leaflets, the anterior and the posterior one, that are attached to a fibrous ring known as the mitral annulus. Its functionality is passively regulated by the pressure gradient between two heart chambers, the left atrium and the left ventricle, with the opening given by the pressure excess of the first compared to the second (during the so-called *diastole* phase), and the closing given in the opposite setting (the *systole* phase). The diseases affecting the MV apparatus can lead to: (1) a narrowing of the valve orifice, thus impairing the flow across the valve during diastole and provoking the so-called *mitral stenosis*; (2) a defect of the coaptation of the MV leaflets during systole, which causes back-flow into the left atrium known as *mitral regurgitation*. This last condition defines the most common cardiac valvular defect and the second most common amenable of surgical intervention [1].

Echocardiography (echo) is a medical imaging technique that

produces 2D and 3D pictures and videos of the heart using ultrasound waves generated by vibrating piezoelectric crystals, scattered and reflected at the biological tissues interfaces, detected and converted by the machine in digital signals. Echo is the standard imaging tool in the clinical routine to perform the diagnosis of most of the heart diseases and dysfunctions, including MV diseases [1–3]. It is inexpensive, non-invasive and it enables both qualitative and quantitative assessment of the myocardium and the MV functions. Clinical practice for the assessment of the MV disease requires physicians to manually trace and measure a plethora of diagnostic parameters. The automatic delineation of the MV annulus and of both the MV leaflets (that we will denote in the following by MV segmentation) could enable physicians to create an automated mechanical model of the MV and to improve the quality of visualization and understanding of the MV pathology.

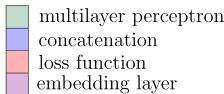
## 1.1. Contribution

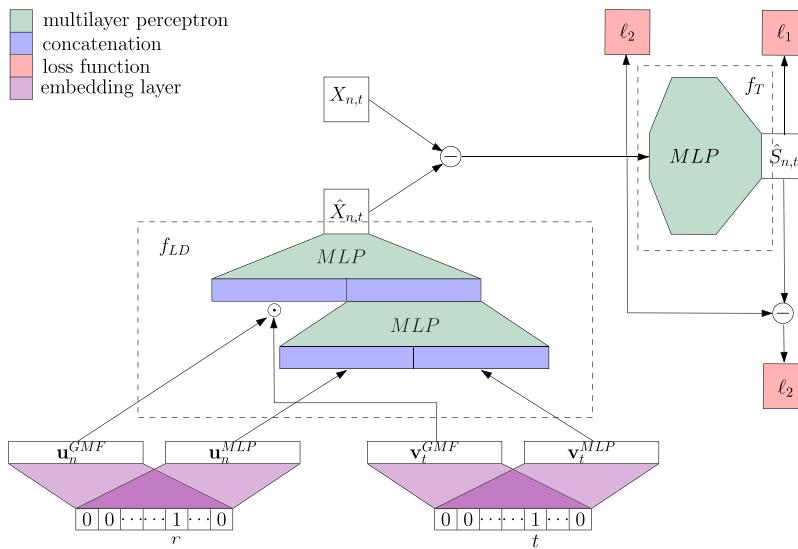
In this paper, we propose NN-MitralSeg<sup>1</sup>, an unsupervised MV segmentation algorithm based on neural collaborative filtering [4,5], that

\* Corresponding author.

E-mail addresses: [luca.corinzia@inf.ethz.ch](mailto:luca.corinzia@inf.ethz.ch) (L. Corinzia), [fabian.laumer@inf.ethz.ch](mailto:fabian.laumer@inf.ethz.ch) (F. Laumer), [alessandro.candreva@usz.ch](mailto:alessandro.candreva@usz.ch) (A. Candreva), [maurizio.taramasso@usz.ch](mailto:maurizio.taramasso@usz.ch) (M. Taramasso), [francesco.maisano@usz.ch](mailto:francesco.maisano@usz.ch) (F. Maisano), [jbuermann@inf.ethz.ch](mailto:jbuermann@inf.ethz.ch) (J.M. Buhmann).

<sup>1</sup> code available at <https://github.com/lucori/NN-MitralSeg>.


 multilayer perceptron  
 concatenation  
 loss function  
 embedding layer



**Fig. 1.** Diagram of the factorization model used in NN-MitralSeg. The network  $f_{LD}$  maps the inputs  $n$  and  $t$  (pixel and frame indices) to the respective latent features  $u_n^{GMF}, v_t^{GMF}, u_n^{MLP}, v_t^{MLP}$  using four different embedding layers (denoted in purple). Such feature vectors are then transformed to the reconstruction  $\hat{X}_{n,t}$ . The vectors  $u_n^{GMF}$  and  $v_t^{GMF}$  are multiplied element-wise and the resulting vector is concatenated (operation denoted in blue) to the output of a MLP (denoted in green) applied to the vectors  $u_n^{MLP}$  and  $v_t^{MLP}$ . Another MLP is then used to produce the approximated signal  $\hat{X}_{n,t}$ . The threshold operator  $f_T$  is then applied to the remainder  $\hat{X}_{n,t} - X_{n,t}$  to give the sparse signal  $\hat{S}_{n,t}$ . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

supports a systematic and fast evaluation of MV health status for medical practitioners. Our method substantially extends our work published in the conference paper [6] and improves on the Robust Non-negative Matrix Factorization method (RNMF), an unsupervised segmentation method proposed in [7] with a three-fold contribution: (i) we use a neural collaborative filtering technique [5,4] that generalizes the matrix factorization and accounts for both linear and non-linear contributions of the myocardial wall motion, in combination with a parametrized threshold operator to learn the high dimensional sparse signal that captures the MV; (ii) we leverage the information of both the optical flow of the sparse signal and of the low dimensional time series representation of the echo to delineate the region of interest (ROI); (iii) we apply post-processing algorithms to improve the final MV segmentation. The method outperforms RNMF on a dataset of 39 patients affected with MV dysfunction and mitral regurgitation, and on an additional independent public dataset of 46 patients extracted from the EchoNet-Dynamic dataset [8]. NN-MitralSeg is further compared to the state-of-the-art supervised segmentation method based on the U-Net neural network architecture [9,10]. U-Net underperforms NN-MitralSeg when trained with up to two annotated frames per videos, while it outperforms it with a higher density of annotation. Nevertheless, its performance on low-quality videos is on par with or worse than NN-MitralSeg at any level of annotation density considered in the study.

## 2. Method

The proposed segmentation model is composed of many stages and follows the structure of other unsupervised methods (see [6,7,11] and the literature review in Section 3). First, the echo video is embedded in a low dimensional space using a factorization technique (e.g. non-negative matrix factorization and variations), then the remainder of the factorization, in the following called sparse signal, is used to delineate the ROI and MV segmentation masks. The next sections present every stage of the algorithm in full details. In this paper, we denote by  $a \in \mathbb{R}$ ,  $a \in \mathbb{R}^d$  and  $A \in \mathbb{R}^{d_1 \times \dots \times d_k}$  respectively a generic scalar, vector, and a  $k$ -rank tensor with  $k \geq 2$ .

### 2.1. Factorization

Here, we present the Neural Matrix Factorization model (NeuMF) used as the first stage of the NN-MitralSeg algorithm.

#### 2.1.1. Model

Each echo is initially represented as a non-negative tensor

$T \in \mathbb{R}_+^{h \times w \times T}$ , where  $h$  and  $w$  are respectively the height and the width of a single frame and  $T$  is the number of frames in the video. We reshape each frame of the echo into a column vector and then concatenate all the columns to get a matrix  $X \in \mathbb{R}_+^{N \times T}$  where  $N = h \cdot w$ . The matrix  $X$  is then embedded in a low dimensional space as follows. We embed each row (pixel)  $n \in N$  and each column (frame)  $t \in T$  into the low dimensional non-negative latent feature vectors  $u_n, v_t \in \mathbb{R}_+^K$ , where  $K$  represents the generalized rank of the factorization model. In the case of linear non-negative matrix factorization the element  $X_{n,t}$  is approximated with the linear product  $u_n \cdot v_t$ . We generalize the simple linear model parameterizing the interaction between the pixel and the frame feature vectors  $u_n$  and  $v_t$  with a feed forward neural network as

$$\hat{X}_{n,t} = f_{LD}(u_n^{GMF} \odot v_t^{GMF}, u_n^{MLP}, v_t^{MLP}; \theta_{LD}) \quad (1)$$

where  $\odot$  is the element-wise product,  $u_n^{GMF}, v_t^{GMF} \in \mathbb{R}^K$  are the generalized matrix factorization (GMF) feature vectors and  $u_n^{MLP}, v_t^{MLP} \in \mathbb{R}^K$  are additional feature vectors that can embed pixel-wise and frame-wise contributions of the entry  $X_{n,t}$  through the multi-layer perceptron (MLP) (see [4])<sup>2</sup>. The function  $f_{LD}$  denotes the *low-dimensional* network with weights  $\theta_{LD}$ . The non-negativity of the latent features is imposed using a non-negative activation function. A diagram of the factorization model is given on the top-left part of Fig. 1. Given the reconstruction  $\hat{X}_{n,t}$ , the difference between  $X_{n,t}$  and  $\hat{X}_{n,t}$  serves as the scalar input to the *threshold network* and is transformed to get the scalar output

$$\hat{S}_{n,t} = f_T(X_{n,t} - \hat{X}_{n,t}; \theta_T). \quad (2)$$

The threshold network is composed by another MLP with weights  $\theta_T$  and non-negative activation function (an illustration of the behavior of this function is given in Fig. 16 and well justifies the name given here). Fig. 1, shows a diagram of the complete architecture.

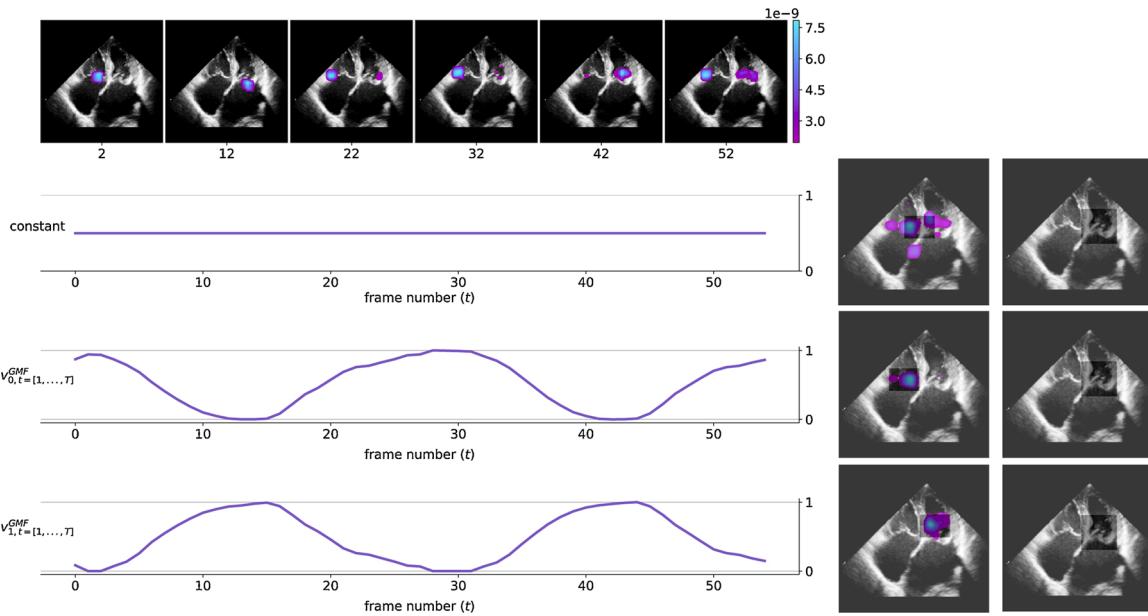
#### 2.1.2. Training

The factorization model is parametrized by the adaptive weights  $\theta_{LD}$ ,

<sup>2</sup> Notice that such model generalizes the  $K$ -rank non-negative matrix factorization given by

$$f_{LD}(u_n^{GMF} \odot v_t^{GMF}, u_n^{MLP}, v_t^{MLP}; \theta_{LD}) = u_n^{GMF} \cdot v_t^{GMF},$$

hence it can reproduce the latter model with a proper choice of the weights of the MLPs that parametrize the function  $f_{LD}$



**Fig. 2.** Improved window detection algorithm with time-masking. The images at the top depict the dense optical flow of  $\hat{S}$  at different time frames while the three plots indicate different time-variant weighting, namely from top to bottom, uniform weighting and the two components of  $v_t^{GMF}$ . In the right two columns, we show the windows selected by the respective time-variant weighting according to Eq. (5), alongside the gold-standard ROI. The uniform (constant) time-variant weighting selects an ROI that is located between the tricuspid and the mitral valve, while the frame embedding components  $v_{0,t}^{GMF}$  and  $v_{1,t}^{GMF}$  capture respectively the tricuspid valve and the mitral valve. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$\theta_T$  and the embedding vectors  $U = \{u_n^{GMF}, u_n^{MLP}\}_{n=1}^N$  and  $V = \{v_t^{GMF}, v_t^{MLP}\}_{t=1}^T$ . The low dimensional network and the embedding layers are trained to ensure that the network produces an accurate approximation of  $X$ . The objective used for reconstruction reads then

$$L_r(\theta_{LD}, U, V) = \|X - f_{LD}(U, V; \theta_{LD})\|_F^2 + \beta \sum_n \|u_n^{GMF}\|_2^2 + \|u_n^{MLP}\|_2^2 + \beta \sum_t \|v_t^{GMF}\|_2^2 + \|v_t^{MLP}\|_2^2 \quad (3)$$

where  $\beta$  denotes a regularization parameter and  $\|\cdot\|_F$  is the Frobenius norm. The optimization of Eq. (3) is performed in two consecutive steps, freezing the embedding vectors  $U, V$  while updating  $\theta_{LD}$ , and then freezing the low dimensional network  $\theta_{LD}$  while updating  $U, V$ .

The threshold network  $f_T$  is applied elementwise on the reconstruction remainder  $X - \hat{X}$  to produce the sparse signal  $\hat{S} = f_T(X - \hat{X}; \theta_T)$ .  $f_T$  is trained to suppress the remainder of the reconstruction, using a  $\ell_1$  penalization (hence imposing sparsity in the sparse signal and enforcing a threshold-like behaviour in the function  $f_T$ ) while reconstructing the signal as much as possible, keeping fixed both  $\theta_{LD}$ ,  $U$  and  $V$ . This goal is achieved by optimizing the loss function

$$L_s(\theta_T) = \|X - \hat{X} - \hat{S}\|_F^2 + \lambda \|\hat{S}\|_1 \quad (4)$$

where  $\lambda$  is the sparsity coefficient and  $\|\cdot\|_1$  denotes the  $\ell_1$ -norm. A summary of the losses used in the model is also given in Fig. 1 by the red boxes, and details on the employed hyperparameters and other details of the optimization routine are specified in Appendix C. The training dynamics is summarized in Section 4.4.

### 2.1.3. Initialization

The initialization of the model parameters is performed in two distinct way: (i) Random initialization (RI), our first method in subsequent experiments, that relies on Xavier initialization [12] of all MLPs parameters  $\theta_{LD}$  and  $\theta_T$  and Gaussian initialization of all the embedding vectors  $U$  and  $V$ ; (ii) matrix-factorization initialization (MFI), that assigns the output of RNMF [7] of the echo video  $X$  with rank  $K$  to the  $K$

dimensional embedding vectors  $u_n^{GMF}, v_t^{GMF}$ . All other parameters are initialized as in RI. The effects of the two initialization schemes are discussed in Section 4.4.

### 2.2. Window detection

After the training of the factorization model, the sparse signal  $\hat{S}$  captures most of the motion expressed by the MV, or by other valves of the heart if they appear in the field-of-view, as well as some speckle noise of the echo; the reader might consult Fig. 3 for a depiction of the sparse signal for the RNMF method in [7] and Fig. 5b for the sparse signal given by our method.

In line with [7,11], we propose a simple window detection (WD) algorithm based on the computation of the norm of the sparse signal  $\hat{S}$  masked by a sliding window. First, the sparse signal  $\hat{S} \in \mathbb{R}_+^{N \times T}$  is reshaped into a 3D array of the same shape as the original video  $\mathbb{R}_+^{h \times w \times T}$ . Denoting by  $\{W_l \in \{0, 1\}^{h \times w}\}_l$  the set of all possible rectangular windows of fixed size  $M$ , we propose a general ROI selection that can be summarized as

$$\begin{aligned} l^* = \arg \max_l \sum_{t=1}^T s_t \|g(\hat{S}_l) \odot W_l\|_2^2 \\ \text{s.t. } \|W_l\|_0 = M \end{aligned} \quad (5)$$

where  $s_t \in \mathbb{R}_+^T$  is a time-variant weight,  $g(\cdot)$  is a generic scalar function applied element-wise on the elements of  $\hat{S}_l$  and  $\|\cdot\|_0$  is the zero-norm that counts the non-zero elements. The selection is made between windows spanning the whole 2D frame, with a fixed stride. We recover the WD used in [7,11] using time-uniform weights and a simple threshold operator for the function  $g$ . This WD method is called TO (for threshold operator) in the following.

In this work, we propose two main variations of this general method that leverage also movement information in the choice of the time weights and of the function  $g$ . The motion of the MV is much faster compared to the myocardium, even when the myocardium appears in the sparse signal. The norm of the dense optical flow [13] can measure

the motion in a video and a large norm is indicative of fast motion. Hence we denoted by OF (for optical flow) the WD method obtained using  $g(\hat{S}_t) = V_{\text{opt}}(\hat{S})_t$  in Eq. (5), where  $V_{\text{opt}}(\cdot)$  is the norm of the dense optical flow. A depiction of the WD algorithms considered so far is given in Fig. 7.

It can be observed that a common source of error for the WD in both TO and OF is the fast movement of other components of the heart, like the tricuspid valve located between the right atrium and right ventricle. Despite being smaller than the MV, this valve can be present in the sparse signal in case of very good quality echos. The window detection used in [7,11] and the one described in Eq. (5) will in general not be able to clearly distinguish between these two valves. We propose an improved window detection algorithm which identifies the motion of different components of the heart during different states within the heart cycle (e.g. the tricuspid and the MV open at different stages, respectively during ventricular diastole and systole). We propose to use the frame embedding vector components  $v_{i,t}^{\text{GMF}}$  for  $i=1, \dots, K$  as time-variant weights in Eq. (5). The optimized ROI selection (denoted by time-masking (TM) in the following) then identifies the right-most window from the  $K$  windows selected by Eq. (5) with the  $K$  different time-variant weights (notice that this step can easily be generalized to other prior assumption on the echo video, like echo view, valve to be selected, etc.). In Fig. 2, the TM method is depicted alongside the behaviour of the frame embeddings  $v_{i,t}^{\text{GMF}}$  for  $K=2$ . Notice that the frame embeddings well describe the periodicity of the heart cycle (as noted already in [7,11]) and that the WD algorithm either selects the tricuspid valve or the mitral valve as ROI depending which component of  $v_{i,t}^{\text{GMF}}$  is used as time-variant weights. An extensive quantitative analysis of the effect of the different WD methods is given in Section 4.5.

### 2.3. Mitral valve segmentation

The segmentation is consequently performed on the pixels of the sparse signal  $\hat{S}_t$  which are above a fixed threshold and enclosed in the ROI, similarly to [7,11] using simple anisotropic 2D diffusion on each frame. We improve this initial segmentation by applying two consecutive post-processing steps: first, we perform erosion and dilation (see [14] for details on morphological operations) on every frame of the sparse signal. In particular, we apply *opening* (dilation of the erosion) to remove noise from the initial segmentation and confine the segmentation mask to the MV. In a second step, we identify all the connected components of the resulting segmentation mask over the 3D volume and discard the small-size components according to a fixed threshold. A qualitative depiction and quantitative analysis of these post-processing steps are also given in Section 4.7.

### 2.4. Overall algorithm

The NN-MitrialSeg algorithm that we propose here is given by composing the NeuMF model for factorization introduced in Section 2.1, the WD method based on the optical flow norm (OF) and time-series masking (TM). The MV is then segmented as described in Section 2.3. An extensive evaluation of the performance of the different algorithms is presented in Section 4.6.

## 3. Related work

MV segmentation in 2D and 3D echo enables automated diagnosis and personalized prognosis of the MV diseases and, therefore, it has received a lot of attention recently. Many early methods are based on active contour algorithms or on other methods that depend extensively on the contribution of human annotators. Active contour algorithms [15,16] require medical practitioners to initialize the segmentation algorithm, placing manually a contour close to the desired position in a given frame [17] or on multiple frames [18]. Then the MV is segmented

on the given frames optimizing a predefined energy function, and the mask is propagated over time with the support of the optical flow [19] or of a dynamical model of the MV [20]. In [21], the proposed method leverages both an active contour algorithm that segments the myocardial walls and a thin tissue detector that finds the valve leaflets. In [22], medical practitioners initialize the segmentation denoting multiple points that are then connected using J-splines.

The first attempts to design a fully automated MV segmentation algorithm are proposed in [7,11]. The 2D echo video is factorized using non-negative 2-rank matrix factorization (NMF) [11] and its robust extension (RNMF) [7].

The 2-rank factorization captures most of the myocardium wall motion, while the high dimensional sparse signal represents the fast MV movement and the echo speckle noise. Then, the MV is segmented using simple diffusion and thresholding of the sparse signal. Despite producing satisfactory results on high-quality echos, these methods perform below clinically acceptable standards on noise perturbed, low-quality videos, mostly due to the misplacement of the ROI of the MV caused by the low expressiveness of the linear model used.

While our segmentation approach is completely unsupervised, supervised methods like those based on neural networks and convolutional architectures [9,23] are rising as the de facto state-of-the-art segmentation method for biomedical 2D imagery matching or exceeding human-level performance. The U-Net architecture [9] emerged as one of the best performing architecture for medical image segmentation and enables clinicians to learn models with very good generalization performance from only a few annotated samples. It has been applied in the context of MV segmentation in [10] and it is here reviewed and tested for comparison.

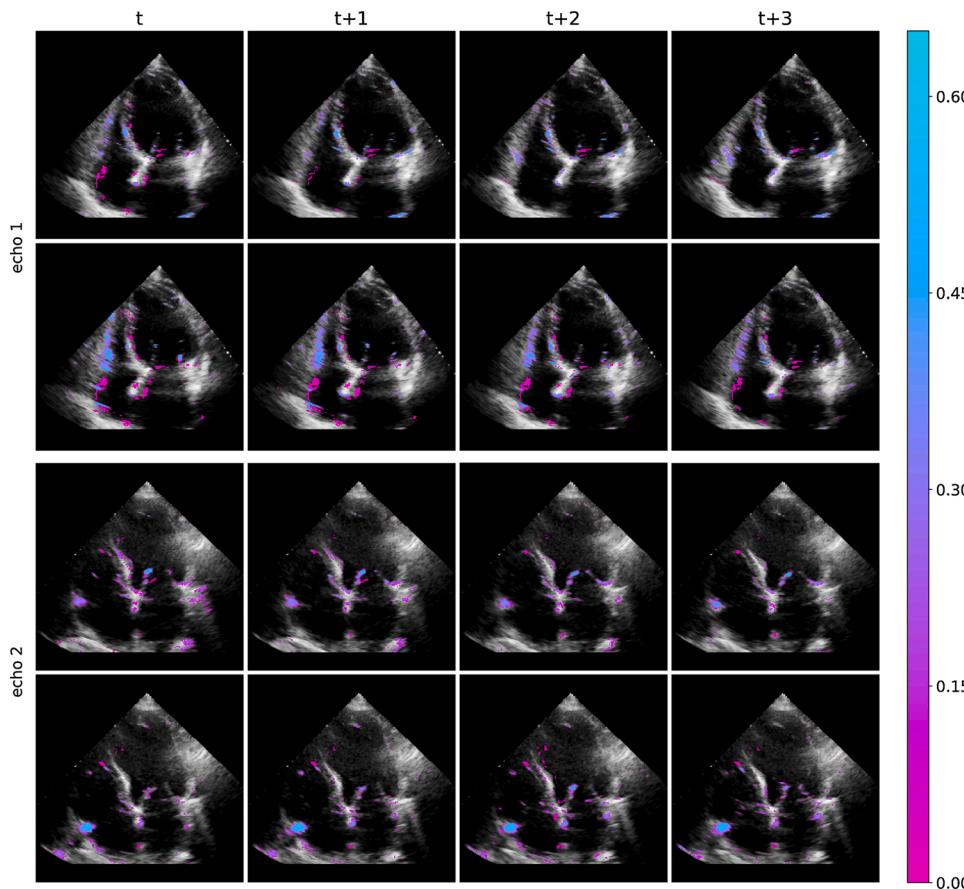
## 4. Experiments and results

### 4.1. Dataset description

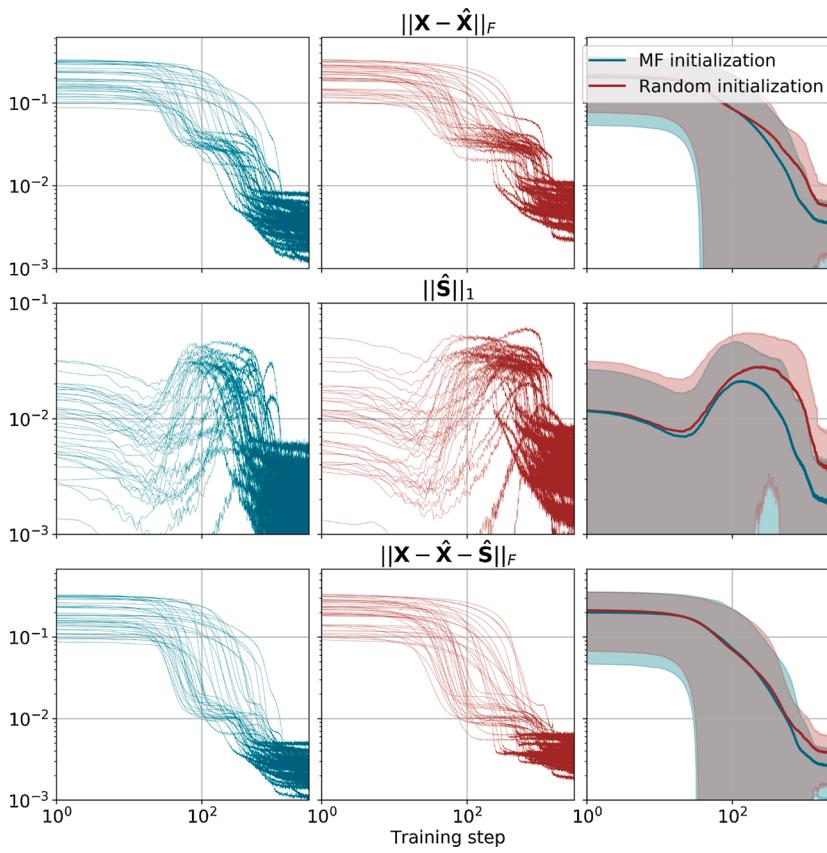
A total of 39 transthoracic echos were obtained from the MitraSwiss Registry, a Swiss-wide prospective registry which includes patients undergoing percutaneous mitral valve repair using the MitraClip system. All patients had moderate-to-severe (3+) or severe (4+) mitral regurgitation of functional or degenerative origin as graded according to current recommendations of the American Society of Echocardiography [24]. Imaging data were processed in an anonymized way and all patients provided written informed consent to be included in the study cohort. Only 4-chamber echo views are used, and for every echo, a rectangular window around the MV (ROI) and three selected frames were annotated by an expert medical doctor. All echos have different  $w$ ,  $h$  and  $T$  dimensions ranging  $w = 600 - 1007$ ,  $h = 579 - 732$ ,  $T = 39 - 159$ . The frame rates vary from 25 Hz up to 60 Hz with an average rate of 43 Hz. The height is first zero-padded to match the size of the width and then the spatial dimensions are down-sampled to a fixed size of  $400 \times 400$  pixels. Additionally, we evaluated the proposed method on another 46 echocardiographic videos from the publicly available EchoNet-Dynamic dataset (see [8] for further details) of the characteristically lower resolution of only  $112 \times 112$  pixels. The performance on this dataset is reported in Table 2 in Appendix A.

### 4.2. Segmentation metrics

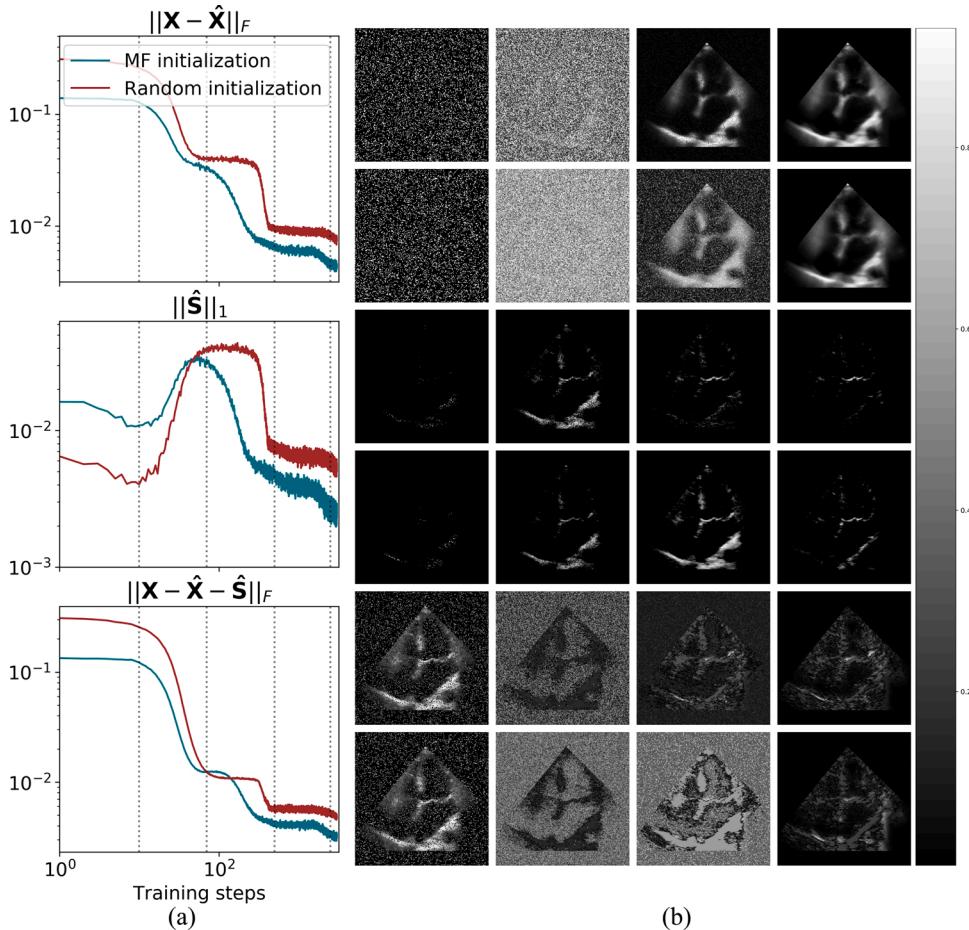
In the following, we denote by  $\mathbf{m}$  and  $\mathbf{m}^{\text{gs}} \in \{0, 1\}^{w \times h}$  two binary masks, with the gold-standard being denoted by  $\mathbf{m}^{\text{gs}}$ . The window detection accuracy  $I$  is defined as the percentage of pixels in the computed ROI that intersect the gold standard mask, formally  $I(\mathbf{m}, \mathbf{m}^{\text{gs}}) = \frac{|\mathbf{m} \cap \mathbf{m}^{\text{gs}}|}{|\mathbf{m}|}$ . Note that in this specific task, the window sizes are fixed and not inferred by the model, hence the accuracy is a reliable measure of performance. Other standard metrics for segmentation tasks are the intersection over Union, that reads instead  $\text{IoU} = \frac{|\mathbf{m} \cap \mathbf{m}^{\text{gs}}|}{|\mathbf{m} \cup \mathbf{m}^{\text{gs}}|}$ , and the



**Fig. 3.** Nonlinear motion of the myocardium captured by the sparse signal  $\hat{S}$  in the RNMF method [7] in two echos. For every echo, four successive frames are reported in the valve closing phase (top row) and opening phase (bottom row). The sparse signal captures the mitral valve movement as well as considerable portions of the myocardium (namely the right ventricle movement for the echo on the top, right atrial movement for the echo at the bottom) in both the opening and closing of the MV. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Fig. 4.** Value of the losses (normalized over the batch size) for all the echos during training. From top to bottom we report respectively the reconstruction loss, the  $\ell_1$  norm of the sparse signal, and the norm of the reminder when also the sparse signal is considered. From left to right we show the curves for MFI, RI and the average over all the echos. In this plot and in the followings, shaded areas encompass a standard deviation interval around the mean. Log scale is used on both axes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Fig. 5.** The same quantities as in Fig. 4 are reported for a single echo showing the characteristic four phases of training indicated by the vertical dashed lines, alongside (b) the depiction of the myocardium, sparse signal, and remainder at the four training steps indicated by the vertical lines. For each plot in (a), the top row corresponds to the MF initialization and the bottom row to the random initialization. All the plots are in log-log scale. Full details in the main text. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Sørensen Dice reads,  $DC = 2 \frac{|m \cap m^{ss}|}{|m| + |m^{ss}|}$ .

#### 4.3. Limitations of RNMF

In Fig. 3, we show two cases of ROI misplacement for the RNMF method [7]. ROI misplacement in the RNMF technique is caused by limitations of the rank-2 matrix factorization, i.e. some of the myocardium movement cannot be separated by a linear decomposition. The mixture of the end-systole and diastole fail to capture some nonlinear movement of the myocardium and this portion of the movement is encoded in the sparse matrix, which essentially corrupts the window detection algorithm. An example of this myocardium movement can be seen in two different echos and two different phases of the cardiac cycle, namely opening and closing of the MV.

#### 4.4. NeuMF training dynamics

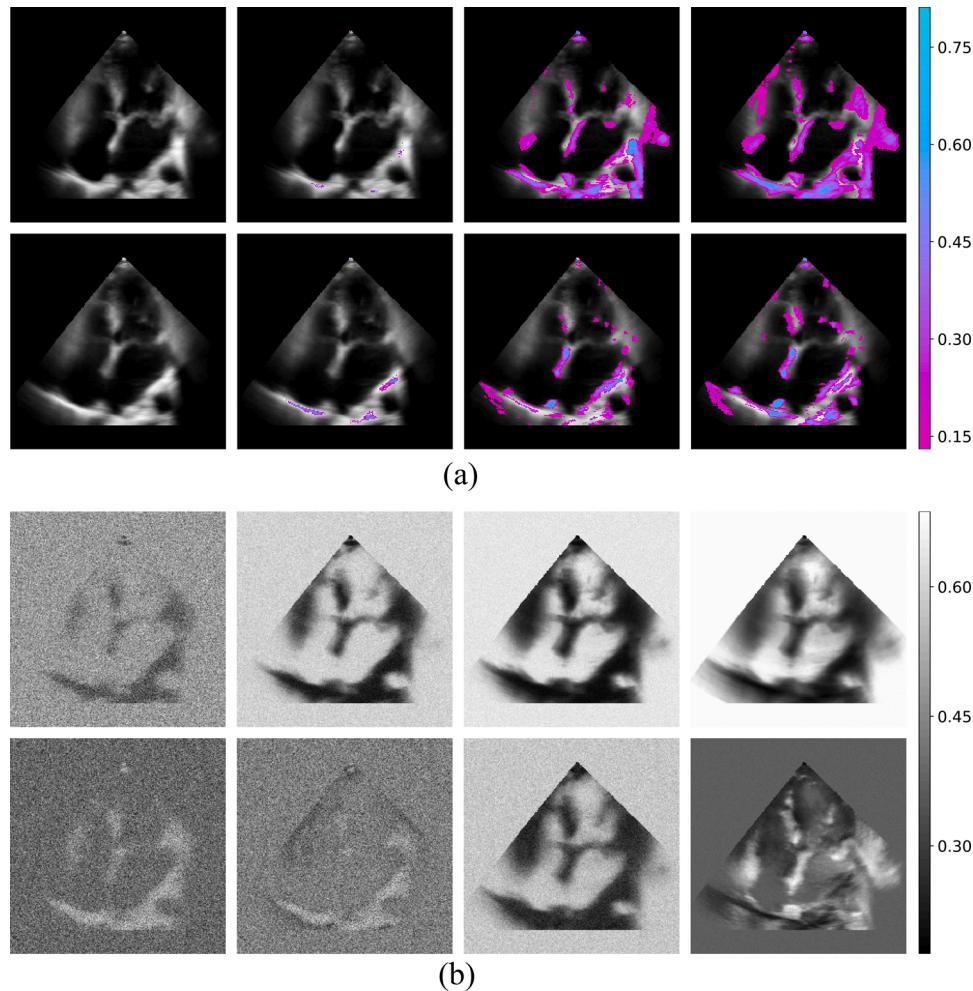
The NeuMF method presented in Section 2.1 shows an interesting dynamics during training, as can be seen in Fig. 4. There, we report the value of the losses  $\ell_{2x} = \|X - \hat{X}\|_F$ ,  $\ell_1 = \|\hat{S}\|_1$  and  $\ell_{2xs} = \|X - \hat{X} - \hat{S}\|_F$  over training for all the echos considered. In Fig. 5, the neural network training dynamics for a single echo is depicted. Four different phases are distinguishable. At the beginning of the training (first dashed vertical line), all quantities are decreasing at a low rate, with the reconstruction  $\hat{X}$  being mostly noise for both RI and MFI as it can be seen in Fig. 5b. Then, we observe a second phase with a steep decrease of both the reconstruction losses  $\ell_{2x}$  and  $\ell_{2xs}$  and the increase of the  $\ell_1$  loss, that is also observed in the images from the second column of Fig. 5b, with the reconstruction and the sparse signal showing spatial structure. After a

plateau of all the quantities, both the reconstruction losses and the sparse signal  $\ell_1$  loss decrease. From the third column of Fig. 5b we can observe that the sparse signal is focusing on the MV region, with the MFI being sparser and with a smaller contribution from the myocardium movement than the RI. This effect can be quantitatively assessed in Fig. 4 where we show the average losses. The MFI reaches better reconstruction in both  $\ell_{2x}$  and  $\ell_{2xs}$  for every echo considered.

As a further analysis of the training dynamics, we monitor the embedding vectors  $u_n^{\text{GMF}}$  for both initialization schemes in Fig. 6a and b. Two main points are noteworthy to observe: (i) For both schemes, the two rows that account for the two dimensions ( $K = 2$ ) of the vectors  $u_n^{\text{GMF}}$  assume different appearances and have both a spatial structure, despite no spatial constraint or information is explicitly embedded in the model (the input is unrolled into a 2D array with spatial structure; a model with explicit spatio-temporal structure encoded by an additional loss is analyzed in Appendix F and its performance are reported in Table 1); hence, the dimension of the vector is exploited to explain the variability of the data. (ii) In the case of MFI in Fig. 6a, the change of the embedding vector, visualized with colours, focuses on the borders of the myocardium walls at the end of the training. This behaviour shows that the complexity of the factorization model manages to express the non-linear motion of the myocardium that is not captured by the linear model. Further embeddings (namely  $v_n^{\text{GMF}}$ ,  $u_n^{\text{MLP}}$ ) are reported in Appendix B for completeness.

#### 4.5. Window detection performance

The performance of the different WD methods considered in Section 2.2 crucially influences the success of mitral valve segmentation. We first observe in Fig. 7 the qualitative behaviour of the RNMF method in



**Fig. 6.** Embedding vectors  $u_n^{\text{GMF}}$  during training, at the four steps indicated by the dashed vertical lines in Fig. 5. The embedding vectors  $u_n^{\text{GMF}} \in \mathbb{R}_+^K$ , with  $n = 1, \dots, w \cdot h$  are reshaped into a  $K \times w \times h$ , and reported as  $K = 2$  slices of shape  $w \times h$  as images. In (a) we report the embedding vector changes during training, superimposed to the embedding at initialization for the MFI. In (b) we report the same vectors for the RI. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 1**

Summary of results of both window detection (WD) and mitral valve (MV) segmentation methods. The second and third columns refer to the WD method, and indicate whether the WD uses thresholding applied to the sparse signal (TO) or its dense optical flow (OF), and whether time masking (TM) is performed. The four centre columns report the performance of the respective method in the WD task, namely from left to right the number of samples with accuracy ( $I$ ) larger than 0.65 and 0.85 ( $I_{65}, I_{85}$ ), average Intersection over Union ( $IoU$ ) and average accuracy. The two rightmost columns report the average performance of the valve segmentation task, according to  $IoU$  and Dice score  $DC$ . The number in parenthesis is the performance obtained with a post-process of the segmentation labels according to Section 2.3. The method introduced in [7] corresponds to the first row (RNMF + TO). NN-MitralSeg corresponds to the last two rows dedicated to the unsupervised methods (NeuMF + OF + TM) with the Gaussian smoothing variation (GS) introduced in Appendix F. We report the U-Net performance dependent on the number of labelled frames per echo available for the training ( $l$ ). The performance on the WD task is not dependent on the number of labels, hence is reported only once. The last two rows describe the performance of the active contour method (see Appendix E for details) dependent on the initialization method (init.) used (given by the naive ROI or its improvement acc.). Since the algorithm requires a manually initialized bounding box we omit reporting the window detection performance.

| Method       | WD | TM  | $I_{65}$ | $I_{85}$ | $IoU$ | $I$              | $IoU$            | $DC$  |
|--------------|----|-----|----------|----------|-------|------------------|------------------|-------|
| RNMF         | TO | ✗   | 35       | 28       | 0.387 | 0.822<br>(0.263) | 0.258<br>(0.395) | 0.390 |
|              | TO | ✓   | 35       | 32       | 0.422 | 0.871<br>(0.277) | 0.270<br>(0.413) | 0.406 |
|              | OF | ✗   | 33       | 27       | 0.383 | 0.811<br>(0.256) | 0.250<br>(0.387) | 0.382 |
|              | OF | ✓   | 36       | 35       | 0.426 | 0.886<br>(0.293) | 0.283<br>(0.424) | 0.424 |
| NeuMF RI     | TO | ✗   | 24       | 13       | 0.275 | 0.612<br>(0.201) | 0.197<br>(0.305) | 0.300 |
|              | TO | ✓   | 25       | 20       | 0.301 | 0.639<br>(0.210) | 0.206<br>(0.316) | 0.310 |
|              | OF | ✗   | 31       | 21       | 0.349 | 0.745<br>(0.230) | 0.229<br>(0.346) | 0.347 |
|              | OF | ✓   | 26       | 20       | 0.331 | 0.694<br>(0.218) | 0.210<br>(0.330) | 0.321 |
| NeuMF MFI    | TO | ✗   | 34       | 30       | 0.384 | 0.816<br>(0.310) | 0.300<br>(0.449) | 0.440 |
|              | TO | ✓   | 36       | 32       | 0.449 | 0.929<br>(0.317) | 0.304<br>(0.462) | 0.451 |
|              | OF | ✗   | 34       | 24       | 0.368 | 0.787<br>(0.271) | 0.262<br>(0.402) | 0.393 |
|              | OF | ✓   | 39       | 33       | 0.453 | 0.939<br>(0.339) | 0.326<br>(0.495) | 0.482 |
| NeuMF MFI GS | OF | ✓   | 35       | 30       | 0.403 | 0.863<br>(0.314) | 0.306<br>(0.461) | 0.447 |
| <i>l</i>     |    |     |          |          |       |                  |                  |       |
| U-Net        | 1  | 38  | 36       | 0.468    | 0.967 | 0.162<br>(0.154) | 0.237<br>(0.222) |       |
|              | 2  |     |          |          |       | 0.300<br>(0.274) | 0.415<br>(0.382) |       |
|              | 3  |     |          |          |       | 0.471<br>(0.458) | 0.615<br>(0.599) |       |
| <i>init</i>  |    |     |          |          |       |                  |                  |       |
| AC           |    | ROI | –        | –        | –     | –                | 0.152            | 0.259 |
|              |    | acc |          |          |       |                  | 0.311            | 0.459 |

four failure cases (rows two and five). We can see that in all cases the failure is caused by strong myocardium movement that is not captured by the linear model, and hence is present in the sparse signal. The NeuMF method alone (row three) decreases the amount of signal involved in the myocardium movement using a higher capacity factorization model, and this leads to mitigate some of the failure cases of RNMF (see Table 1 for an extensive quantitative evaluation). The WD

methods based on the optical flow (OF) are presented in rows five to seven and show a sharper signal of moving pixel in the sparse signal. We can see that this mitigates most of the RNMF failures, giving always a stronger signal in the MV area, alongside however other areas like those capturing the tricuspid valve and myocardium movements. These different contributions are separated by the TM method that is reported at the bottom row, leading to a satisfactory ROI delineation in these four echos considered.

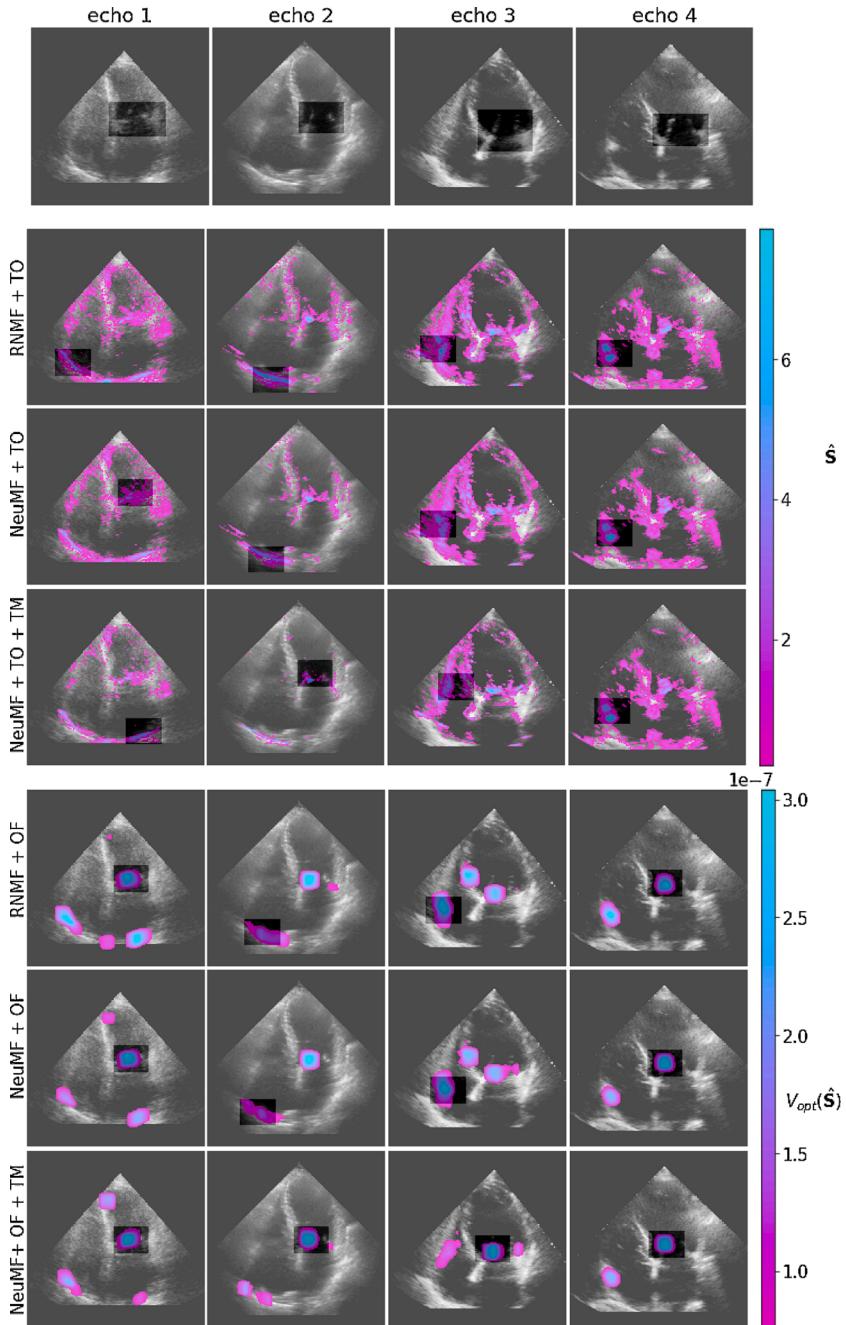
The quantitative assessment is performed in different stages. In Table 1, we report the number of WD success cases ( $I_{65}$  and  $I_{85}$ ) for two accuracy thresholds (namely 0.65 and 0.85 that represent respectively a level of satisfactory and good WD) and the average  $IoU$  and accuracy  $I$  for the different methods distinguished by the factorization method (RNMF, NeuMF with RI and MFI), the type of window detection function used (threshold operator TO or optical flow OF) and whether TM has been applied. We observe that NN-MitralSeg outperforms all other methods, which includes the state-of-the-art RNMF as described in [7], in all the metrics considered, except for  $I_{85}$ . Interestingly, applying either NeuMF or the optical flow WD individually in the MV segmentation framework does not increase the WD performance relative to RNMF. The low expressiveness of the linear model in RNMF leaves a strong presence of the myocardium movement in the sparse signal. The dense optical flow is then computed on a sparse signal with a mild amount of myocardium movement and hence its performance decreases as it can be seen in row one and three of Table 1 and in Fig. 7. On the contrary, NeuMF has a high expressiveness and captures most of the myocardium movement in the echo. However, the high expressiveness also captures a small amount of the MV movement, hence the WD method based on the original sparse signal has a lower performance in NeuMF compared to RNMF. The combination in NN-MitralSeg uses the benefits of NeuMF, optical flow and TM to achieve better performance. The high expressiveness of NeuMF reduces the brightness of the pixels in the MV region in  $\hat{S}$  but captures a large portion of the myocardium movement. The dense optical flow calculation on  $\hat{S}$  is not corrupted by the myocardium movement, resulting in the best performance in window detection among all methods. The random initialization method is also reported in Table 1 and shows a substantial gap in performance compared to the MFI method.

The effect of the TM method is reported in Fig. 8. The TM method increases the performance of the WD method with a statistically significant margin and notably uniformly over all the echos (it increases the performance on more than 90% of the echos in all four WD method considered).

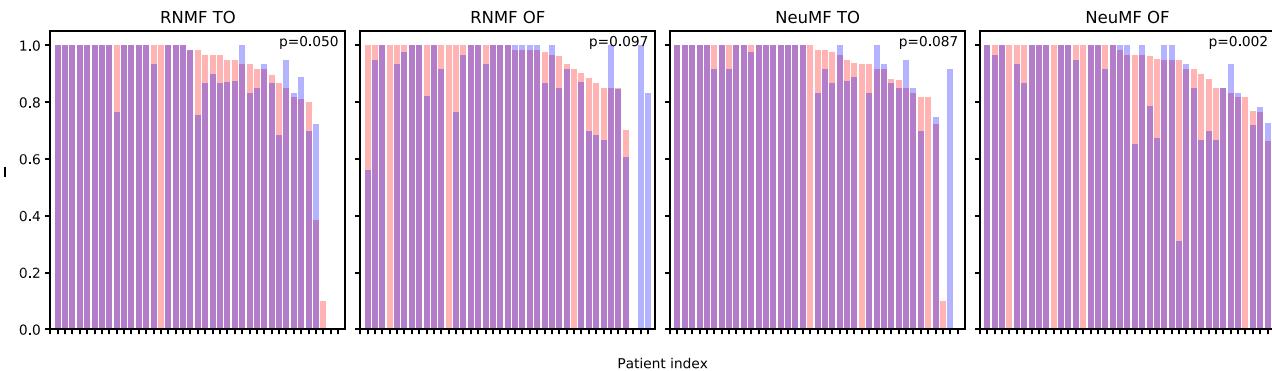
For completeness, we also report in Table 1 the performance of two other segmentation methods, namely, the state-of-the-art supervised deep learning method based on the U-Net architecture [9,10], and an active contour method [25] which requires manual initialization. The U-Net, despite being a supervised method and hence not comparable with the other methods considered so far, provides a strong benchmark for the task. It can be seen that on average it outperforms the NN-MitralSeg. It is however less robust and highly influenced by the echo quality, underperforming on the  $I_{65}$  score and low-quality video as it can be seen in Fig. 10 (two plots on the right). Full details on the implementation of the U-Net method is given in Appendix D. More details about the active contour method are presented Appendix E.

#### 4.6. Mitral valve segmentation performance

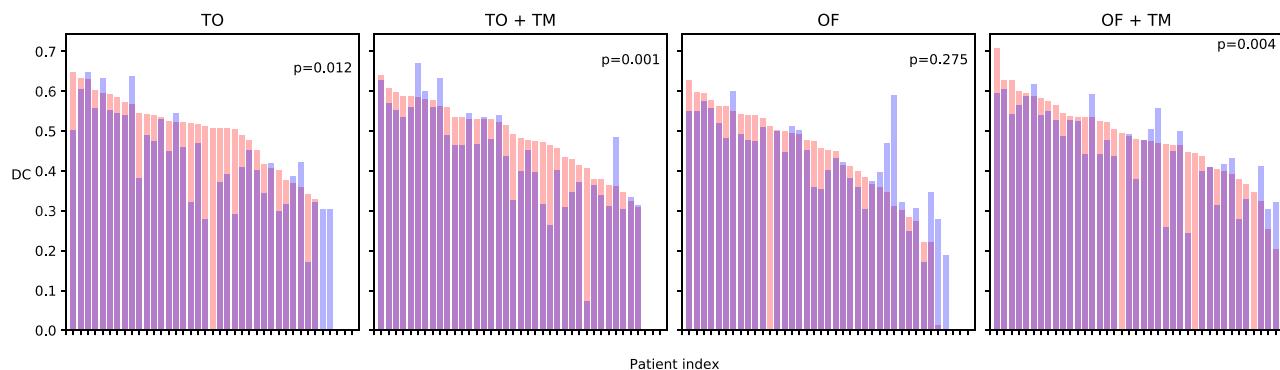
The MV segmentation performance is assessed here using  $IoU$  and the dice coefficient  $DC$  with respect to the gold standard available as introduced in Section 4.1. In Table 1, we report both scores for all the methods considered here. We can observe the same trends as for the WD task, with a distinction being the effect of the NeuMF, that increases the performance in the MV segmentation by a statistically significant margin. For some videos, the RNMF manages to place the ROI more



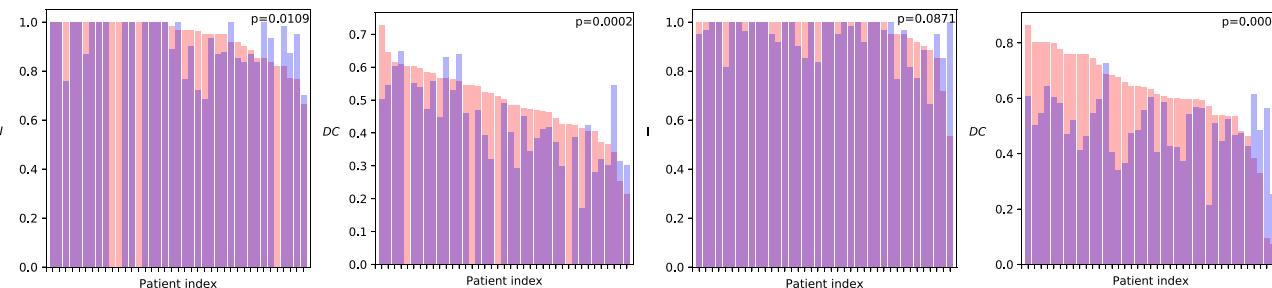
**Fig. 7.** Window detection algorithm outputs on four different echos. All the ROIs are indicated as the shaded area in each frame. The top row corresponds to the gold-standard, and every column represents one single echo. Rows from two to four represent respectively the algorithms RNMF + TO, NeuMF + TO, NeuMF + TO + TM, rows from five to seven represent respectively the algorithms RNMF + OF, NeuMF + OF, NeuMF + OF + TM. Full details in the main text. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Fig. 8.** Effect of the time masking method on the performance of different WD methods for each video. The title of each plot reports the factorization method and the processing of the sparse signal used (TO or OF) on which the TM is applied. Red and blue bars represent respectively the method with and without TM. Accuracy  $I$  is sorted according to the method that uses TM, and on the top right of each plot, the  $p$ -value for a one-sided  $t$ -test is reported. Better viewed in colours. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Fig. 9.** Effect of the NeuMF on the performance of different segmentation methods reported as in Fig. 8. Red and blue bars are respectively NeuMF and RNMF based methods. The  $DC$  score is sorted according to the NeuMF factorization. On the top right of each plot, the  $p$ -value for a one-sided  $t$ -test is reported. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



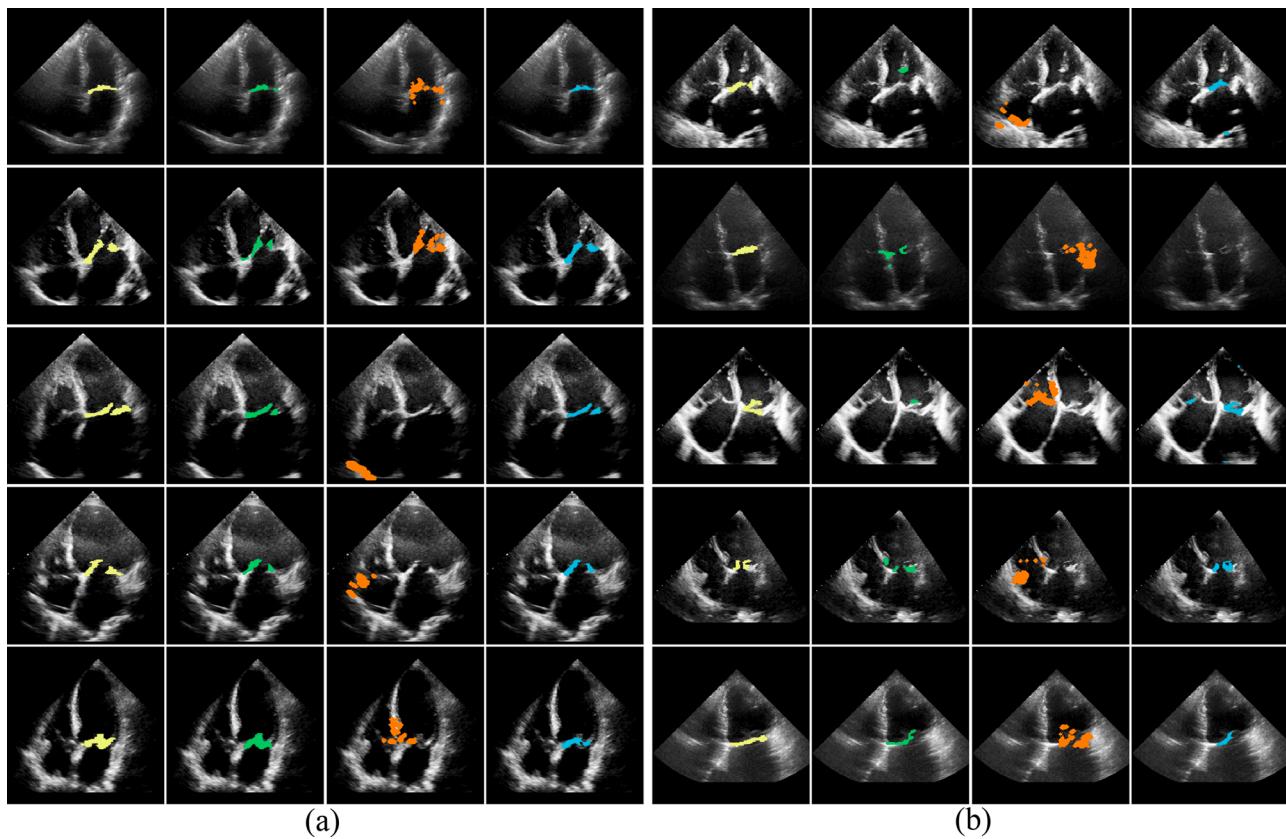
**Fig. 10.** Performance on both the WD and MV segmentation tasks for the NN-MitralSeg, RNMF [7] and U-Net [10] reported as in Fig. 8. The first two plots report the performance for NN-MitralSeg (red) and RNMF (blue) from left to right on WD and MV segmentation (scores sorted according to NN-MitralSeg). The last two columns report the performance for U-Net (red) and NN-MitralSeg (blue) from left to right on WD and MV segmentation (scores sorted according to U-Net). On the top right of each plot, the  $p$ -value for a one-sided  $t$ -test is reported. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

accurately than the WD based on NeuMF factorization. Even for these videos, the segmentation performance of NeuMF exceeds the RNMF method. This effect is documented in Fig. 9 and it is related to the sparse signal  $\hat{S}$  that is more focused on the MV area than for the RNMF based methods. The sparse signal of NeuMF is also sparser on average than the one from the competing techniques, resulting in a higher MV segmentation performance on all echos with a satisfactory WD accuracy, and hence a higher MV segmentation performance on average.

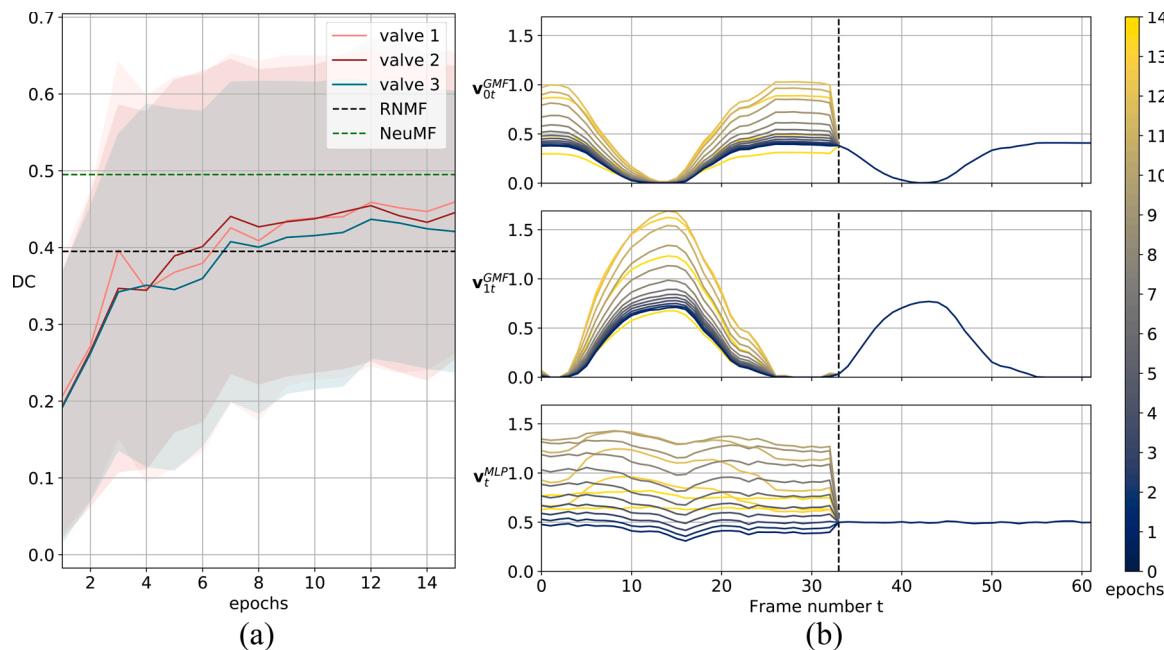
In Table 1, we also observe that the supervised approach on average outperforms the NN-MitralSeg algorithm on IoU/Dice for the mitral valve segmentation with a high level of annotation. However, U-Net

performance is highly dependent on the number of available training samples. With only two labels per echo, the supervised approach drops in segmentation quality below the NN-MitralSeg performance. Using only one labelled frame per echo impairs the predicted segmentation even in a stronger way. A detailed comparison of the NN-MitralSeg and the state-of-the-art of both unsupervised [7] and supervised [9,10] is provided in Fig. 10.

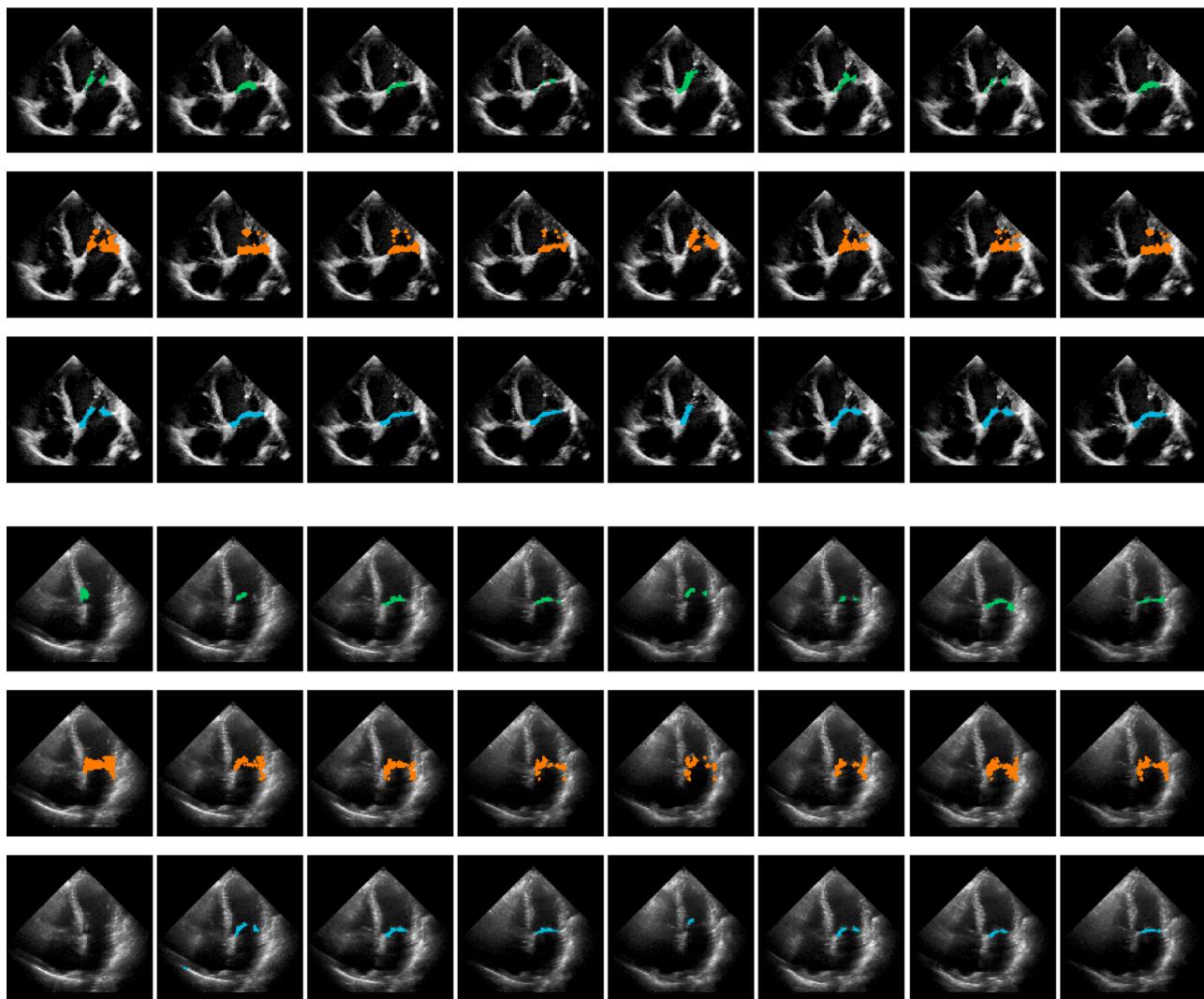
A detailed comparison of the MV segmentation produced by NN-MitralSeg and the state-of-the-art RNMF [7] and U-Net [9] is documented in Fig. 11a and b and in Fig. 13 where we show respectively the segmentation masks and the gold standard for the highest and lowest



**Fig. 11.** The MV segmentation masks for the five echos with the (a) highest and (b) lowest Dice coefficients according to NN-MitralSeg. From left to right: ground truth (yellow), NN-MitralSeg (green) RNMF (red) and U-Net (blue). The WD algorithm for the RNMF approach is significantly less reliable than the one used in NN-MitralSeg. The U-Net performs better than the other method on average but fails on low-quality echos (second row, (b)) or if the tricuspid valve is clearly visible (third row (b)). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Fig. 12.** Experiment of the deployment of the pretrained NN-MitralSeg method on new echos. Full details in the main text. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)



**Fig. 13.** The MV segmentation mask on two different echos (top three rows and bottom three rows) for eight frames separated by five time-steps each. NN-MitraSeg (green), R-NNMF (orange) and U-Net (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

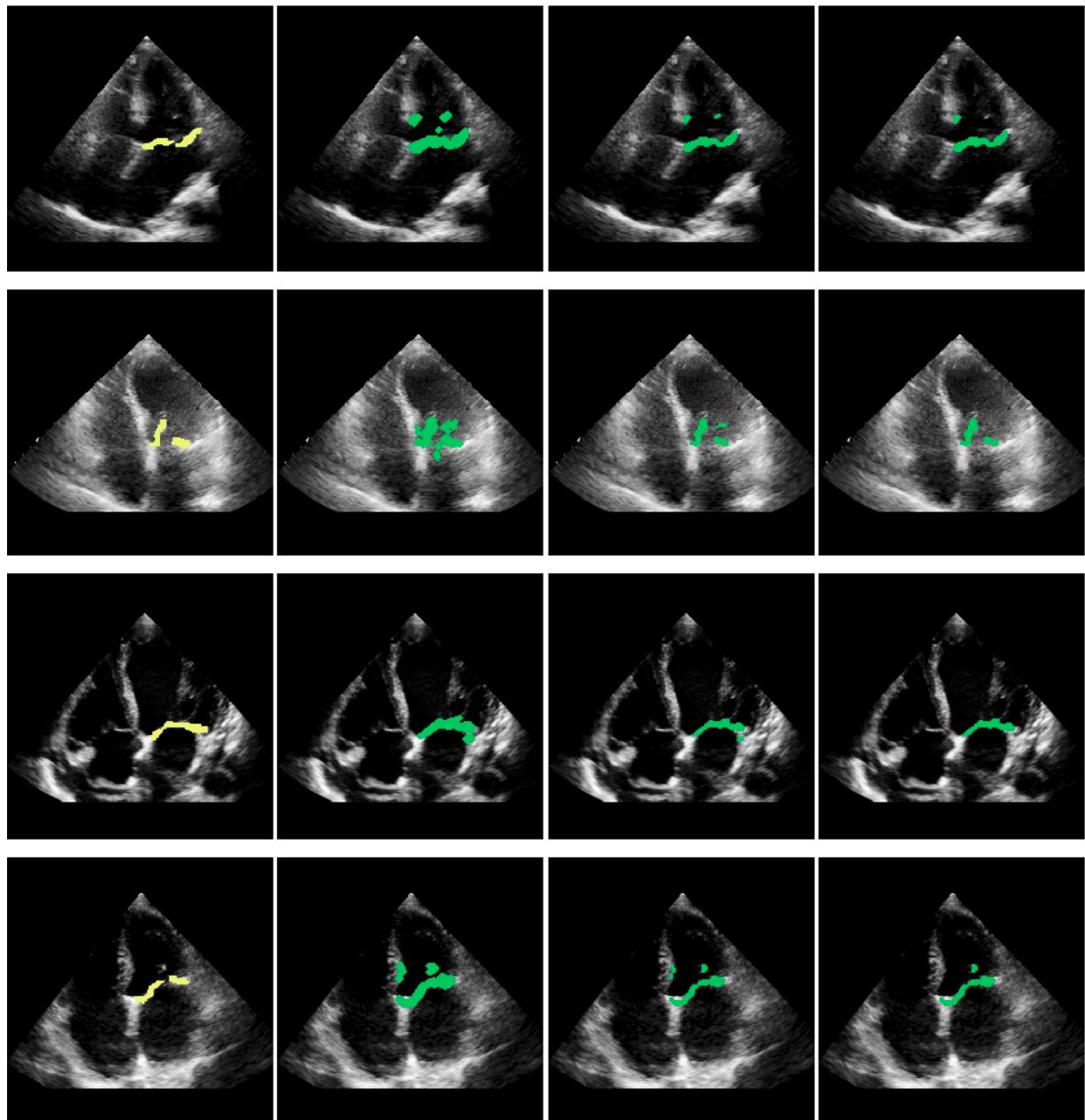
five scoring echos (according to our method), and the time consistency of the MV segmentation masks produced by the different methods.

#### 4.7. Post-processing

The effect of the post-processing steps used in the NN-MitralSeg algorithm is qualitatively depicted in Fig. 14 and quantitatively reported in Table 1. Applying erosion and dilation confines the initial segmentation closer to the MV. Consequently, isolated parts of the segmentation masks are detected as the smallest connected components calculated over the complete echo and hence removed. Post-processing applied to the RNMF and NeuMF algorithms leads to better IoU and Dice scores. The IoU improves on average by 2.6% for RNMF, by 2.1% for NeuMF (RI) and by 3.8% for NeuMF (MFI). The Dice coefficient increases on average by 1.7% for RNMF, by 1.5% for NeuMF (RI) and by 2.4% for NeuMF (MFI). The predicted segmentation masks by the U-Net are not improved by post-processing (see again Table 1, last three rows). The U-Net does not capture other fast-moving parts within the ROI, and segments only the parts belonging to the mitral valve. Applying morphological operations, i.e. erosion, reduces the size of the predicted area and hence decreases the segmentation performance.

#### 4.8. Computation issues and real-world deployment

The method so far described, together with all the other unsupervised methods based on low dimensional factorization method and subsequent outlier detection, requires the training of the full model (neural network and spatial-temporal embeddings) on every new echocardiography video available. In our experiments, this task required an average of  $13 \pm 8$  min per video on a single GeForce GTX TITAN X GPU. Despite this computational burden being limited, it can result in a barrier in the clinical practice, especially in time-critical scenario like in intensive care units or operative procedure. For this reason, in the following, we design an experiment that shows the viability of the method in these circumstances. The echos are split into two consecutive parts such that the two parts contain respectively two and one labelled MVs. The NN-MitralSeg model is then trained only on the first part of the echo (with two labelled MVs) and tested on the second part. This setting simulates the scenario in which an echo has been processed by the NN-MitralSeg method, and subsequently the method is deployed on a new echo of the same patient, hence the neural network weights and all the spatial embeddings can be redeployed with no further training. In Fig. 12b, we can observe the qualitative behaviour of the time-variant embedding during training, with the embedding vector partitioned into a trained and a frozen section. The performance of the method is



**Fig. 14.** Effect of the post-processing applied used in the NN-MitralSeg: The gold standard (yellow) and the segmentation (green) is depicted for four different echos. Columns from left to right report the valve segmentation mask respectively with no post-processing (second column), with the application of morphological operations (erosion and dilation, third column), and with the removal of the smallest connected components (calculated over the whole echo, last column). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

reported in Fig. 12a, with the individual dice scores for the three valves, averaged on the whole dataset. We can observe that despite the model being trained only on one part of the echo, the segmentation performance on the last valve increases during training, hence the model can generalize well to the section of echos that has not been trained on. We can further observe that despite the model performing on average worse on the last valve than on the other two, it can outperform the RNMF

method (dashed black line) trained on the full echos. The drop in the dice coefficient from the fully trained model (green dashed line) is due to the decreased amount of trained data and amounts in 10% decrease on average.

## 5. Conclusion and future work

We proposed NN-MitralSeg, a fully automated and unsupervised mitral valve segmentation algorithm based on non-linear matrix factorization using neural network collaborative filtering. An echocardiography video is decomposed into a low dimensional signal that captures the linear and non-linear myocardial wall motion, and a high dimensional sparse signal that accounts for the echocardiography noise and mitral valve movement. The mitral valve is then segmented from the sparse signal using thresholding, diffusion algorithms and morphological operations. This method outperforms the state-of-the-art fully automated unsupervised algorithm in a data-set of 39 videos with patients suffering various mitral valve dysfunctions and in an independent public data-set, in both the task of positioning the rectangular region of interest and in the accuracy of the dense mitral valve mask. NN-MitralSeg also compares favourably with the state-of-the-art supervised method, (i) outperforming it on a low level of annotation and (ii) being more robust to low-quality echo at every level on annotation considered. These performance benefits demonstrate that prior assumptions of the structure of echocardiography videos are of great importance in the tasks considered. Furthermore, while NN-MitralSeg

could easily be extended to 2CH or 3CH echo views, the supervised approach would most likely fail if not retrained again with new labels.

Despite the modest size of the dataset, the reported experiments contain echo videos with a larger variability than a healthy control dataset of the same size, thereby documenting the robustness of the method. Possible future developments include the use of both sparse ground truth segmentation masks and dense (inaccurate) annotation generated by unsupervised algorithms (like NN-MitralSeg) to train deep networks for segmentation in a weakly-supervised learning scenario [26]. To ensure relevance in clinical practice, the segmentation algorithms should also be extended to efficient online processing. Enforcing a factorization-like structure in the embedding space of the U-Net architecture can be expected to retain the best of both approaches. This design choice would provide practitioners with segmentation algorithms that could be deployed in real-time echocardiography during mitral valve intraoperative procedures.

## Acknowledgments

Joachim M. Buhmann has been supported by PHRT(ETH) for the SWISSHEART Failure Network.

## Appendix A. Results EchoNet-Dynamic dataset

We evaluated our segmentation algorithm on the publicly available echocardiographic video dataset (EchoNet-Dynamic Dataset [8]). 46 videos were selected and for each video three frames were labelled by medical experts. The resolution of the videos is only  $112 \times 112$  pixels. Despite the low resolution, the NN-MitralSeg manages to achieve good results (see Table 2). However, due to the low resolution of the videos, using optical flow (OF) for window detection impairs the performance. In Table 2, we compare our algorithm against the previous state of the art unsupervised segmentation algorithm based on robust non-negative matrix factorization (RNMF) [7] and against the supervised U-Net [9,10].

**Table 2**

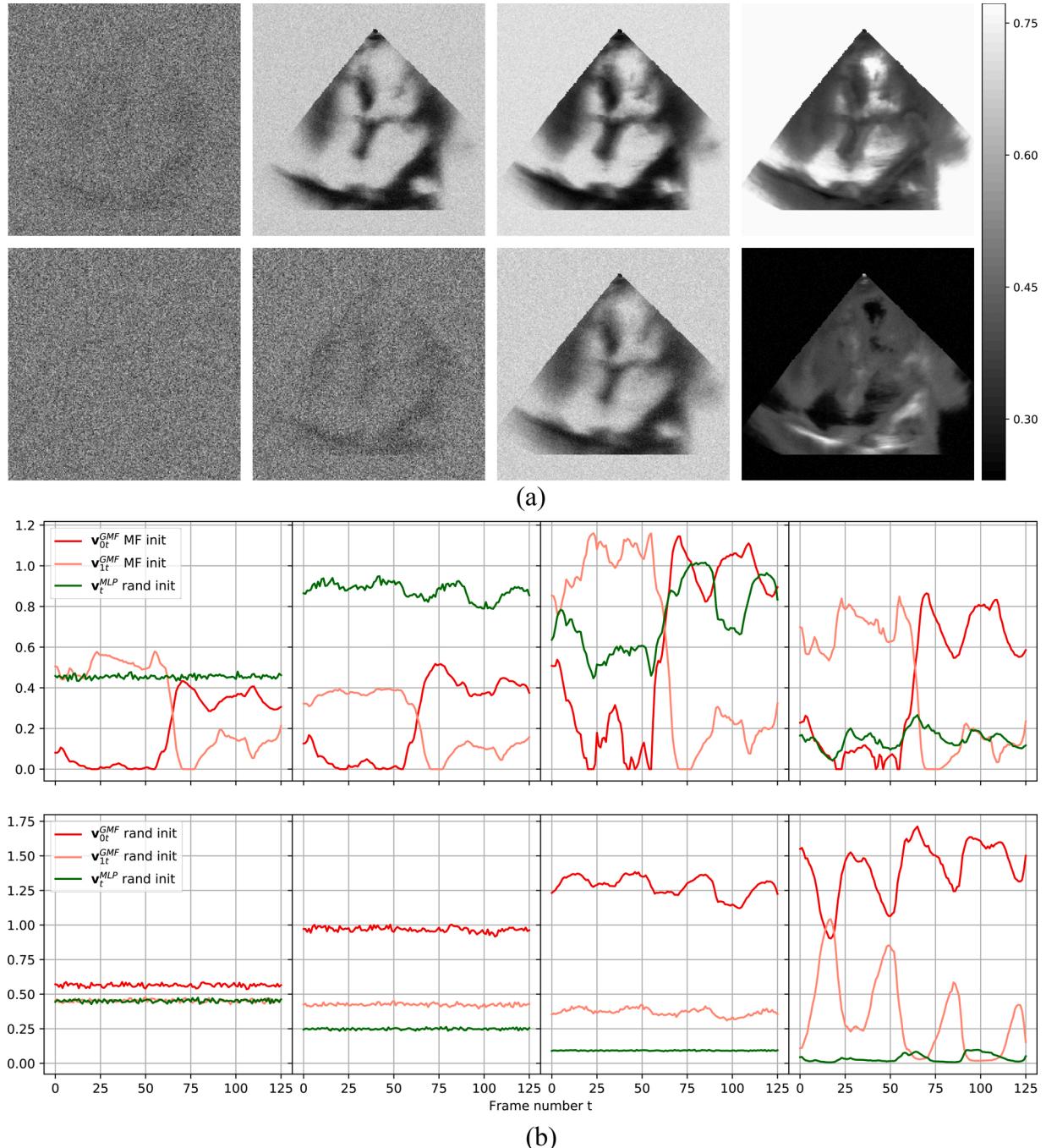
Summary of the results for both window detection and mitral valve segmentation methods evaluated on the EchoNet-Dynamic dataset. For the unsupervised method RNMF and NeuMF, the post-process refinement introduced in Section 4.7 is used. Full details in the caption of Table 1.

| Method    | WD | TM | $I_{65}$ | $I_{85}$ | $IoU$ | $I$   | $IoU$ | $DC$  |
|-----------|----|----|----------|----------|-------|-------|-------|-------|
| RNMF      | TO | ✗  | 36       | 30       | 0.367 | 0.790 | 0.197 | 0.314 |
|           | TO | ✗  | 41       | 40       | 0.422 | 0.887 | 0.321 | 0.461 |
| NeuMF MFI | TO | ✓  | 45       | 42       | 0.451 | 0.957 | 0.375 | 0.531 |
|           | OF | ✗  | 23       | 12       | 0.229 | 0.517 | 0.159 | 0.239 |
|           | OF | ✓  | 29       | 16       | 0.298 | 0.668 | 0.202 | 0.302 |
| $\bar{I}$ |    |    |          |          |       |       |       |       |
| U-Net     |    | 3  | 46       | 41       | 0.447 | 0.946 | 0.515 | 0.660 |

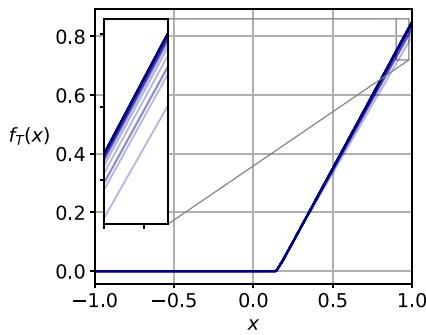
## Appendix B. Additional visualizations

In Fig. 15a, we report the embedding vector  $v_t^{\text{MLP}}$  for the two different NeuMF initialization (RI and MFI) on one echo in four sequential stages of training. We can observe that the model embeds a frame that resembles an average frame, being more blurred and less defined on the myocardium walls, and hence accounting for a bias term of the GMF column of the model. We can also notice that the MFI is faster in building a spatial structure in the embedding vector.

In Fig. 15b, we can observe the time-variant embedding vectors  $v_t^{\text{GMF}}$  and  $v_t^{\text{MLP}}$ . Note that the MLP embedding vector  $v_t^{\text{MLP}}$  is randomly initialized in both cases, and it gains a complex time-structure in the MFI while being close to zero and with little time structure in the RI. In Fig. 16 we also report the thresholding function values learned by the network over all echos considered. The thresholding function changes only slightly between different echos and behaves indeed as a threshold operator.



**Fig. 15.** Embedding vectors during training, for both MFI (top) and RI (bottom) at the four step indicated by the dashed vertical lines in Fig. 4. In (a) the images are built as in Fig. 6a. For both cases the embedding dimensions are  $K = 2$  and  $K' = 1$ .



**Fig. 16.** Threshold network values in the interval  $\{-1, 1\}$  and the end of the training for all the echos considered.

### Appendix C. Model specifications for NeuMF

Across both datasets (MitraSwiss registry and EchoNet) and all echos, the hyperparameters of the model are kept fixed as follow. The Adam optimizer [27] is used with a learning rate of 0.005 and a batch size of 10,000. The dimensions of the latent features are kept constant across all echos at  $K = 2$ , and  $K' = 1$ . The sparsity coefficient and regularization parameter are also kept constant across all videos at  $\lambda = 0.3$  and  $\beta = 0.1$ . The window size used for the window detection procedure is fixed to  $60 \times 80$ . The standard deviation of the Gaussian to smooth the temporal derivatives in the optical flow algorithm is set to 3.5. Both networks (low dimensional and threshold) consists of three fully connected (FC) layers with 10 units each. The latent features and the first two FC layers have ReLU and the last FC layer has sigmoid activation to give  $\hat{X}_{n,t} \in [0, 1]$ . The softplus activation function is used to obtain non-negative embedding vectors. Xavier initialization is performed with gain one, while all the Gaussian initializations are performed on the embedding vectors with average 0.5 and standard deviation 0.01. The number of training epochs is held constant among all echos and is fixed to 15.

### Appendix D. U-Net implementation

The network is fully convolutional and consists in total of 18 convolutional layers with a kernel size of  $3 \times 3$  and ReLU activation function, and a  $1 \times 1$  convolutional layer with a sigmoid activation function for the output layer. The encoder consists of four max pooling operations ( $2 \times 2$ ) which are applied after every second convolutional layer. In the decoder part, up-convolutions (transposed convolutions) with a kernel size of  $2 \times 2$  and strides of  $(2, 2)$  are used. For the mitral valve segmentation and the ROI selection, the same U-Net architecture described above is used. The Adam optimizer [27] is used with a learning rate of 0.001,  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . The binary cross-entropy loss<sup>3</sup> is used for training. A batch-size of 16 is used for both the mitral valve segmentation and the ROI selection. The models are trained with early stopping with the validation loss evaluated on 10% of the training. The models' performance are evaluated with 13-fold cross-validation. The splitting is performed over videos (36 videos for training and 3 videos for testing) for every cross-validation fold, and the average of the testing performance is reported in both Table 1 and Table 2. For the MV segmentation either one, two or three labelled frames per video are used, while for the ROI prediction one label for each frame (one ground-truth ROI per video) is available. The final mitral segmentation is achieved by thresholding the output of the networks at 0.5. The final ROI selection is determined by averaging the ROI predictions of individual frames of one echo, calculating the centre point of the averaged prediction and then taking a fixed window size of  $80 \times 60$  pixel around this centre point. This is done to ensure a fair comparison with the NN-MitralSeg window detection algorithm.

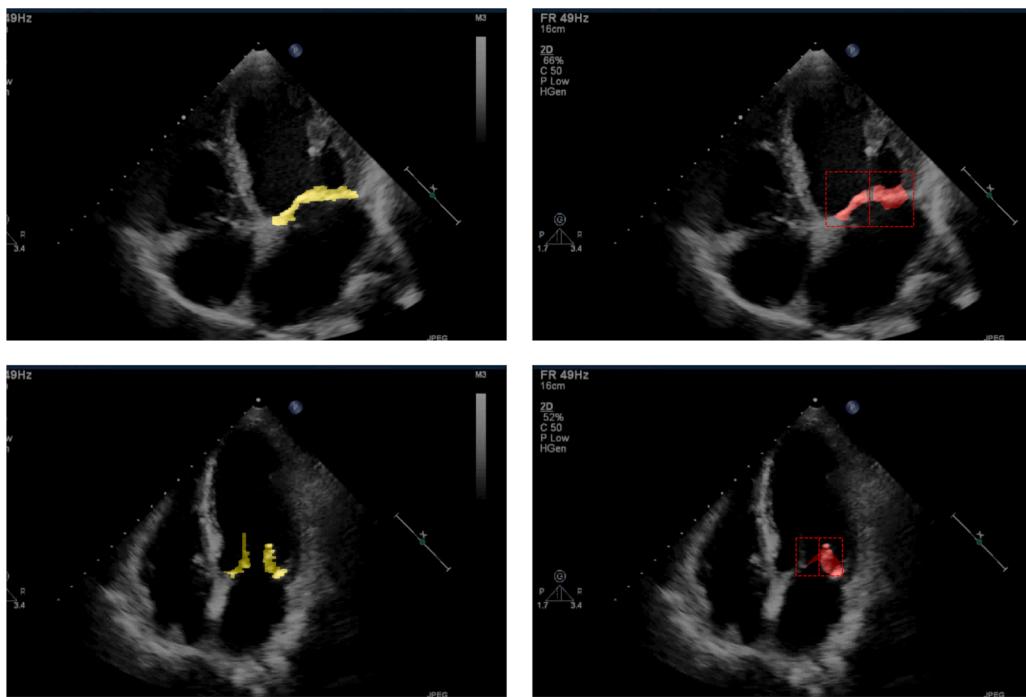
### Appendix E. Segmentation with the active contour method

For the implementation of the active contour method (AC) [25] we use the scikit-image library for image processing [28]. The AC method uses a spline to define the contour of the MV to be segmented and minimizes an energy functional that is in part defined by the image (so that the spline follows contours) and part by the spline's shape, i.e. length and smoothness (so that the splines cannot be arbitrary rugged). Active contour methods require good manual initialization to achieve acceptable performance. In case the bounding box used for the initialization is too large the method fails to segment the mitral valve accurately. The main reason is that the myocardium close to the MV provides sharper intensity gradients than the MV itself. In Fig. 17 we provide two selected examples of the predicted segmentation of the AC against the ground truth label in case of success (top row) and failure (bottom row) of the method.

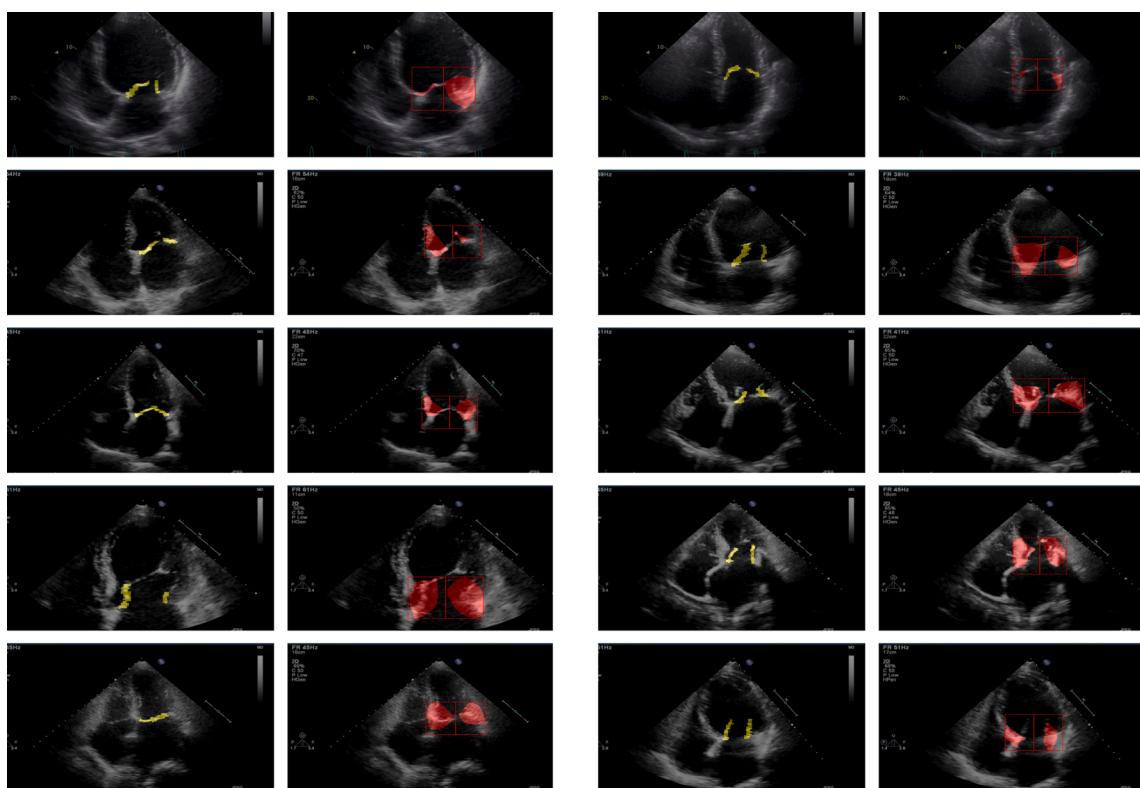
To have a fair comparison with the other methods, we use as the spline initialization the ROI given for every video. In Fig. 18a we depict a random selection of segmentation predictions and in Table 1 and Table 2 (second last row) we report the quantitative performances. It can be observed that the results are particularly poor, with the AC underperforming the unsupervised RNMF benchmark despite using the ROI information. A further improvement is given in Fig. 18b (with quantitative evaluation given again in Table 1, last row), where a more accurate bounding box is computed using prior knowledge of the MV leaflets shapes and sizes.

The heavy supervision in the form of the initialization improvement gives an important boost to the performance of the AC method (+0.2 in Dice score, Table 1). Nonetheless, the AC (with naive an accurate initialization) underperforms both unsupervised and more recent supervised methods (based on deep learning), hence can be considered outdated for the task considered here.

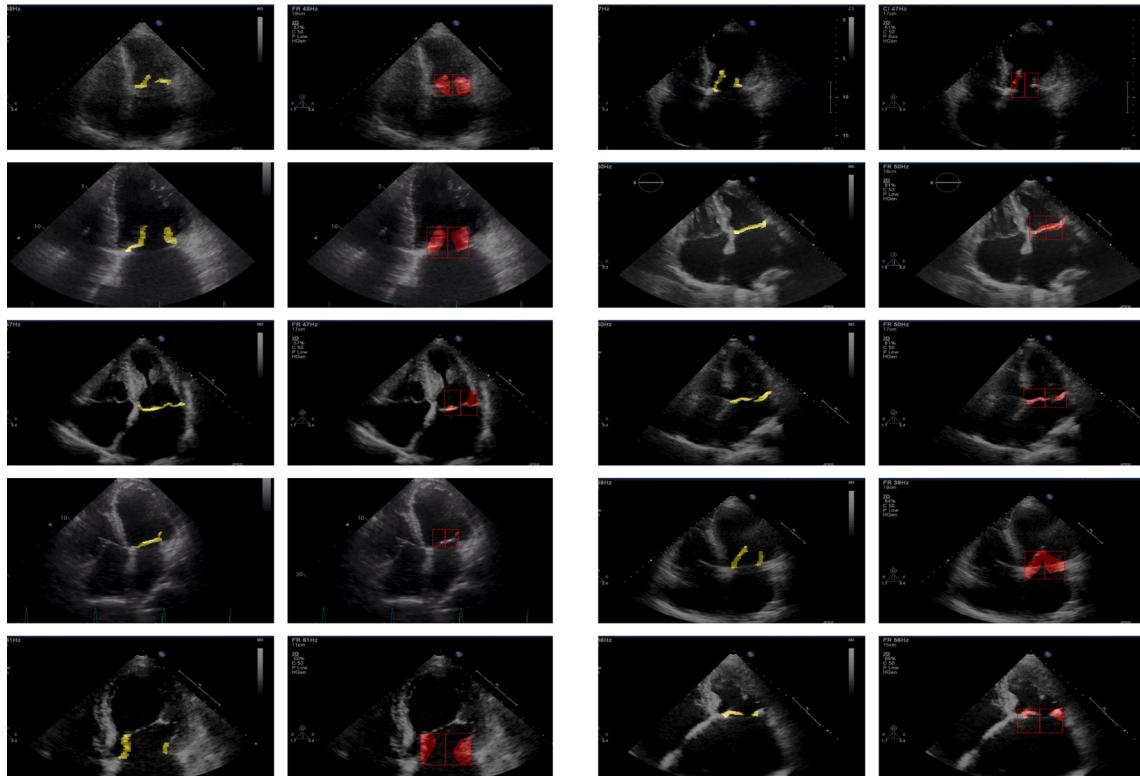
<sup>3</sup> Applying weighted binary cross-entropy to account for the highly imbalanced class distribution, i.e. valve vs. background, did not lead to better results.



**Fig. 17.** Top row: Example of the active contour method success in delineating the mitral-valve. Bottom row: Example where the active contour method fails to delineate one of the leaflets due to a bright spot in the other (right) leaflet (yellow: ground-truth, red: AC prediction). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

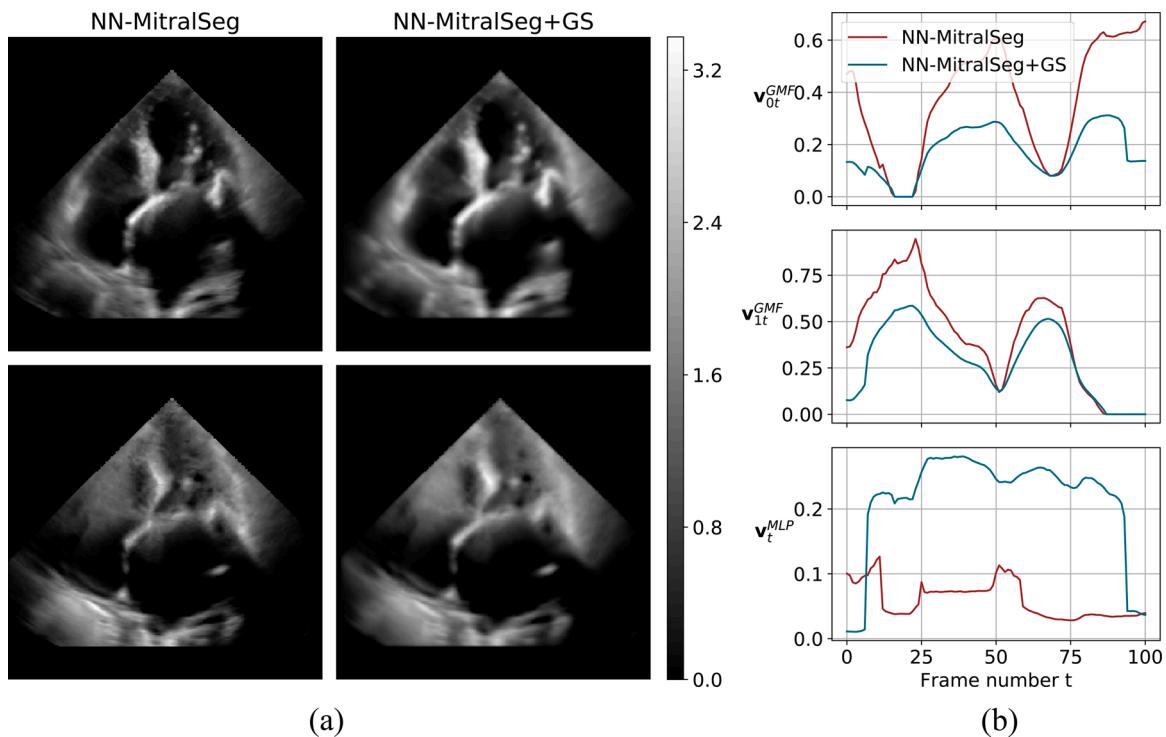


(a)



(b)

**Fig. 18.** Examples of the mitral valve segmentation obtained by the active contour method (yellow: ground-truth, red: AC prediction). (a) Initialization given by the ROI provided by medical experts, (b) Improved initialization. Full details in the main text. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Fig. 19.** Comparison of the embedding vectors for the NN-MitralSeg method and the same method with an additional kernel convolutional loss on both the spatial (a) and temporal (b) domain. The embeddings are obtained in both cases at the end of the training.

#### Appendix F. Gaussian smoothing loss

In order to encode spatio-temporal information in the NeuMF model we here design an additive loss that acts on the the embedding vectors. Let us reshape the collection of embeddings  $\{u_n^{GMF}\}_{n=1}^N$  with  $u_n^{GMF} \in \mathbb{R}_+^K$  into a 3D array  $U_{3D}^{GMF} \in \mathbb{R}_+^{w \times h \times K}$  of the same shape of single frame, with an additional dimension (the channel dimension) given by  $K$ . The same reformatting can be applied also to the other embedding vectors obtaining  $U_{3D}^{MLP} \in \mathbb{R}_+^{w \times h \times K'}$ ,  $V_{2D}^{GMF} \in \mathbb{R}_+^{T \times K}$  and  $V_{2D}^{MLP} \in \mathbb{R}_+^{T \times K'}$ . Given this reformatting, we can add a loss function that penalizes spatial and temporal variations of the embedding as

$$L_{gs} = \| U_{3D}^{GMF} - \text{Ker}_{3D} * U_{3D}^{GMF} \|_F^2 + \| U_{3D}^{MLP} - \text{Ker}_{3D} * U_{3D}^{MLP} \|_F^2 + \| V_{2D}^{GMF} - \text{Ker}_{2D} * V_{2D}^{GMF} \|_F^2 + \| V_{2D}^{MLP} - \text{Ker}_{2D} * V_{2D}^{MLP} \|_F^2 \quad (6)$$

where  $\text{Ker}_{3D}$  and  $\text{Ker}_{2D}$  are respectively a 3D kernel (spatial kernel + channel dimension) and a 2D kernel (temporal kernel + channel dimension). In the experiment, we used Gaussian kernels with a kernel size of 15 and variance fixed to 1, identical for all the channels, and we compare the effect of the additive loss on the NN-MitralSeg method (NeuMF factorization, with temporal masking and optical flow window detection). In Fig. 19, we can observe the qualitative effect of the  $L_{gs}$  loss, that has a smoothing effect on both the spatial (Fig. 19a) and temporal embeddings (Fig. 19b). We can further observe in Fig. 19b how the loss  $L_{gs}$  penalizes also the magnitude of the embedding hence acting as an additional  $\ell_2$  regularization, with the embeddings  $v^{GMF}$  obtained with Gaussian smoothing loss being consistently smaller than the original ones. Quantitative results are given in Table 1, where we can observe that the spatio-temporal continuity loss decreases the performance in all metrics, suggesting that the additional constraint impairs the expressiveness of the model.

#### References

- [1] Hayek E, Gring CN, Griffin BP. Mitral valve prolapse. *Lancet* 2005;365(9458): 507–18.
- [2] Zamorano J, Cordeiro P, Sugeng L, Perez de Isla L, Weinert L, Macaya C, et al. Real-time three-dimensional echocardiography for rheumatic mitral valve stenosis evaluation: an accurate and novel approach. *J Am College Cardiol* 2004;43(11): 2091–6.
- [3] Baumgartner H, Hung J, Bermejo J, Chambers JB, Evangelista A, Griffin BP, et al. Echocardiographic assessment of valve stenosis: eae/ase recommendations for clinical practice. *J Am Soc Echocardiogr* 2009;22(1):1–23.
- [4] He X, Liao L, Zhang H, Nie L, Hu X, Chua T-S. Neural collaborative filtering. *Proceedings of the 26th international conference on world wide web* 2017:173–82.
- [5] Dziugaite GK, Roy DM. Neural network matrix factorization. 2015 (arXiv preprint), arXiv:1511.06443.
- [6] Corinzia L, Provost J, Candreva A, Tamarasso M, Maisano F, Buhmann JM. Unsupervised mitral valve segmentation in echocardiography with neural network matrix factorization. Conference on artificial intelligence in medicine in Europe 2019:410–9.
- [7] Dukler Y, Ge Y, Qian Y, Yamamoto S, Yuan B, Zhao L, et al. Automatic valve segmentation in cardiac ultrasound time series data. *Medical imaging 2018: image processing*, vol. 10574 2018:105741Y.
- [8] Ouyang D, He B, Ghorbani A, Lungren MP, Ashley EA, Liang DH, et al. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. 2019.
- [9] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *International conference on medical image computing and computer-assisted intervention* 2015:234–41.
- [10] Costa E, Martins N, Sultan MS, Veiga D, Ferreira M, Mattos S, et al. Mitral valve leaflets segmentation in echocardiography using convolutional neural networks. In: 2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG); 2019. p. 1–4.
- [11] Zhou X, Yang C, Yu W. Automatic mitral leaflet tracking in echocardiography by outlier detection in the low-rank representation. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR); 2012. p. 972–9.

- [12] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth international conference on artificial intelligence and statistics 2010:249–56.
- [13] Farnebäck G. Two-frame motion estimation based on polynomial expansion. Proceedings of the 13th scandinavian conference on image analysis, SCIA'03. Berlin, Heidelberg: Springer-Verlag; 2003. p. 363–70.
- [14] Serra J, Soille P. Mathematical morphology and its applications to image processing, vol. 2. Springer Science & Business Media; 2012.
- [15] Isard M, Blake A. Contour tracking by stochastic propagation of conditional density. European conference on computer vision 1996:343–56.
- [16] Blake A, Isard M. Active contours: the application of techniques from graphics, vision, control theory and statistics to visual tracking of shapes in motion. Springer Science & Business Media; 2012.
- [17] Mikic I, Krucinski S, Thomas JD. Segmentation and tracking of mitral valve leaflets in echocardiographic sequences: active contours guided by optical flow estimates. Medical imaging 1996: image processing, vol. 2710 1996:311–21.
- [18] Shang Y, Yang X, Zhu L, Deklerck R, Nyssen E. Region competition based active contour for medical object extraction. Comput Med Imaging Graph 2008;32(2):109–17.
- [19] Mikic I, Krucinski S, Thomas JD. Segmentation and tracking in echocardiographic sequences: active contours guided by optical flow estimates. IEEE Trans Med Imaging 1998;17(2):274–84.
- [20] Schneider RJ, Tenenholz NA, Perrin DP, Marx GR, Pedro J, Howe RD. Patient-specific mitral leaflet segmentation from 4d ultrasound. International conference on medical image computing and computer-assisted intervention 2011:520–7.
- [21] Burlina P, Sprouse C, DeMenthon D, Jorstad A, Juang R, Contijoch F, et al. Patient-specific modeling and analysis of the mitral valve using 3d-tee. International conference on information processing in computer-assisted interventions 2010: 135–46.
- [22] Siefert AW, Icenogle DA, Rabbah J-PM, Saikrishnan N, Rossignac J, Lerakis S, et al. Accuracy of a mitral valve segmentation method using j-splines for real-time 3d echocardiography data. Ann Biomed Eng 2013;41(6):1258–68.
- [23] Hariharan B, Arbeláez P, Girshick R, Malik J. Hypercolumns for object segmentation and fine-grained localization. Proceedings of the IEEE conference on computer vision and pattern recognition 2015:447–56.
- [24] Zoghbi WA, Enriquez-Sarano M, Foster E, Grayburn PA, Kraft CD, Levine RA, et al. Recommendations for evaluation of the severity of native valvular regurgitation with two-dimensional and doppler echocardiography. J Am Soc Echocardiogr 2003;16(7):777–802.
- [25] Kass M, Witkin A, Terzopoulos D. Snakes: active contour models. Int J Comput Vis 1988;1(4):321–31.
- [26] Zhou Z-H. A brief introduction to weakly supervised learning. Natl Sci Rev 2017;5 (1):44–53.
- [27] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014 (arXiv preprint), arXiv:1412.6980.
- [28] Van der Walt S, Schöller J, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in python. PeerJ 2014;2:e453.