APSIPA Transactions on Signal and Information Processing, 2023, 12, e39

Original Paper

Optical Flow Regularization of Implicit Neural Representations for Video Frame Interpolation

Weihao Zhuang[*,†], Tristan Hascoet[†], Xunquan Chen, Ryoichi Takashima and Tetsuya Takiguchi

Kobe University, Kobe 657-8501, Japan

ABSTRACT

Recent works have shown the ability of Implicit Neural Representations (INR) to carry meaningful representations of signal derivatives. In this work, we leverage this property to perform Video Frame Interpolation (VFI) by explicitly constraining the derivatives of the INR to satisfy the optical flow constraint equation. We achieve state-of-the-art VFI on Adobe-240FPS, X4K and UCF101 datasets using only a target video and its optical flow, without learning the interpolation operator from additional training data. We also found that constraining the INR derivatives not only enhances the interpolation of intermediate frames but also improves the ability of narrow networks to fit observed frames. By limiting the INR derivatives, we were able to improve the network's efficiency in fitting observed frames, which could lead to more advanced video compression techniques and optimized INR representations. Our work highlights the potential of Implicit Neural Representations in video processing tasks and provides valuable insights into their utilization for signal derivatives.

[*] Corresponding author: Weihao Zhuang, zhuangweihao@stu.kobe-u.ac.jp

[†] Equal contribution. The order of appearance was decided by the toss of a coin.

Received 03 March 2023; Revised 13 August 2023

ISSN 2048-7703; DOI 10.1561/116.00000218

1 Introduction

Signal processing often involves core concepts defined by continuous functions and their derivatives. For instance, surfaces manifest as continuous manifolds in space, and motion represents spatial change over time. However, modern digital hardware fundamentally operates on a discrete level: digital sensors take discrete, regularly-timed observations of the world; computers store and manipulate discrete signal representations.

Classical methods try to portray continuous concepts on discrete signal representations [1, 2, 7], relying on simplified assumptions. These often involved constant first or second derivatives of the signal between sequential observations. The limited applicability of these handcrafted heuristics, along with the increasing accuracy of Machine Learning (ML) methods, has driven a widespread adoption of ML in recent signal processing research. ML methods use large datasets to derive signal statistics instead of using handcrafted heuristics.

Shifting to computer vision, Video Frame Interpolation (VFI) exemplifies such developments. VFI models strive to interpolate intermediate frames between sequential video frames. Most successful methods use optical flow to guide pixel intensity interpolation from the observed frames' pixel grid to the intermediate frames' pixel grid. Traditional methods assume constant motion or acceleration fields between sequential frames. Each pixel's value in the interpolated intermediate frame is determined by shifting pixel intensities of observed frames along optical flow directions, then interpolating these shifted intensities onto the intermediate frame's pixel grid. However, such methods face two main limitations:

- The optical flow is prone to errors due to occlusions, external illumination variations, etc.
- Assumptions of constant motion field or its derivatives do not often hold true in practice.

Both limitations arise from the same issue: discretization. The constant brightness assumption, which informs the optical flow, and the constant motion field assumption used in interpolation, only hold true on an infinitesimal scale. These time deltas are usually much smaller than those used in practical Frames Per Second (FPS).

ML approaches [8, 11, 12, 19, 20] propose to learn the frame interpolation operator from large video collections, rather than making explicit assumptions about the signal. While these methods have achieved impressive benchmark performance, they are susceptible to generalization errors due to domain shifts. In fact, discrepancies between the training set distribution (such as

VFI benchmark videos) and the target video distribution can affect ML model performance. This can be due to differences in motion range, exposure time, FPS, and blur [30].

Meanwhile, research on implicit representations is working towards better discrete representations of continuous signals. In recent years, Implicit Neural Representations (INR) have demonstrated several advantages over explicit representations. This involves representing signals as Neural Networks (NN), with early successes in 3D shape representations [17]. Of particular interest is the work of SIREN [25], which has shown that representing images with Multi Layer Perceptrons (MLP) with sine activation functions allows for meaningful representations of signal derivatives.

Inspired by this work, we question if SIREN could guide the interpolation process of VFI by controlling the exact derivatives of the signal. This approach avoids the traditional method's discretization pitfalls by examining the finite differences between sequential discrete frames. We achieve this by ensuring the derivatives of SIREN representations satisfy the optical flow constraint equation. Specifically, they must be orthogonal to the video's optical flow, which we calculate using current state-of-the-art OF models.

We found this method outperforms most existing ML-based approaches on small motion range benchmarks. This was achieved without relying on ML for the interpolation operator. In this sense, our method resembles classical VFI approaches, but applies the optical flow constraint on the INR's exact gradient rather than wrapping the OF on discrete explicit frame representations. Our method is thus not affected by any discrepancies between training and test data.

Moreover, our approach can sample any number of frames between observed frames due to the continuous nature of the representation. Besides its application to VFI, we also show that constraining the model's gradient enhances the ability of narrow MLPs to fit the signal. This suggests potential applications in INR optimization and video compression.

To summarize the contributions of this work, we show that:

- SIREN representations of videos can be constrained so as to satisfy the OF constraint in their exact derivatives.

- Such representations reach state of the art VFI on Adobe-240FPS, X4K and UCF101 datasets, without learning a residual flow nor interpolation operator.

- The OF constraint not only allows SIREN to generate intermediate frames, but also improve the ability of narrow SIREN to fit observed frames.

On the other hand, our approach, in its current form, presents several limitations:

- Optimization of the INR is time-consuming, which hinders our ability to work on full resolution videos for time constraints.
- Our method currently only works on limited motion ranges; it does not match state of the art ML models on large motion ranges.

In light of these limitations, the aim of this paper is not to provide a standalone, production-ready VFI system. Instead, we aim to present actionable insights on a simple method that can be either built upon or integrated into existing models.

The remainder of this paper is organized as follows: We briefly present some related work in Section 2, the detail of our method in Section 3, and design several experiments to highlight the merits of our approach in Section 4. Finally, we discuss current limitations and present potential ways to address them in Section 5, before concluding in Section 6.

2 Related Work

Our research lies at the intersection between video frame interpolation and video scene representation as we apply INR of videos to the problem of VFI. In this section, we review related works on these two topics.

Implicit Neural Representations have met early success in shape representation and 3D rendering [16–18]. Since then, a number of studies have attempted to apply INR to different signals including audio [10, 25], images [5, 6], videos [3, 22, 23], medical imaging and climate data [6]. In [25], the authors have shown that MLP with sine activations could fit representations of images with meaningful representations of their gradient, and that such models could be optimized to satisfy constraints on their gradients. Together, these two findings have motivated our idea to apply the optical flow constraint to the gradient of SIREN representations of videos. A series of recent studies have applied INR to video compression [3, 31], with some works [3] even reporting higher PSNR than practical codecs on high compression rates. Although closely related to video compression, we differ from these works as we focus on VFI. Most related to ours is the concurrent work by [23], which also uses INR for VFI. Their approach, CURE, differs from ours in scope: they propose to learn a prior on the INR, while we only focus on leveraging INR to guide the interpolation process using a given optical flow.

Video Frame Interpolation research has largely relied on optical flow to guide the video frame interpolation process [1, 2, 7]. Most studies have assumed uniform optical flow between consecutive frames so as to linearly interpolate intermediate frames along the optical flow directions. One exception is the work of [27], in which the authors propose to take into account acceleration to perform the interpolation, leading to quadratic interpolation. Our work constrains only the first derivatives of the signal. We differ from classical works in that we apply the OF to the exact representation derivatives, so that we do not need to assume constancy of signal derivatives on any time interval. Recent OF-based VFI research leverages deep learning for optical flow estimation and interpolation. Super-SloMo [8] is an important study of such methods. The authors use a deep learning model to predict the forward and backward flows of intermediate frames, and warp the two surrounding frames to obtain the intermediate frames. RRIN [12] uses residual learning to optimize the performance of [8] at the motion estimation bound. AMBE [20], a current state-of-the-art VFI method, proposes an asymmetric motion estimation method based on [19], which enhances the quality of interpolated frames by loosening the linear motion constraint. Kernel-based approaches such as AdaCof [11] avoid explicit separation of motion estimation and wrapping stages and instead directly interpolate intermediate frames from consecutive observed ones.

3 Method

We consider a ground-truth video as a continuous signal v mapping continuous spatial (x, y) and temporal (t) coordinates to RGB values:

$$\begin{aligned} v &: (x, y, t) \rightarrow (R, G, B) \\ v &: \mathbb{R}^3 \rightarrow \mathbb{R}^3 \end{aligned} \tag{1}$$

Our goal is to find a continuous function f_θ, parameterized by $\theta \in \Theta$, with a minimum distance d to the ground-truth signal:

$$\begin{aligned} f_\theta &: (x, y, t) \rightarrow (R, G, B) \\ s.t.\ \theta = &min_\Theta \iiint d(f_\theta(x, y, t), v(x, y, t)) dxdydt \end{aligned} \tag{2}$$

where the distance function d may either be the Peak Signal to Noise Ratio (PSNR) or the Structural Similarity Index Measure (SSIM). To do so, we only have access to regularly sampled observations of the signal v (i.e. the explicit representation of the video), which we denote as:

$$\begin{aligned} &\mathcal{V} \in \mathbb{R}^{T \times H \times W \times 3} \\ &s.t.\ \mathcal{V}_{xyt} = v(x, y, t) \end{aligned} \tag{3}$$

where T represents the number of frames in the video, and $H \times W$ represents the spatial resolution. We use SIREN as the parameterized function f_θ. The most straightforward way to approximate Equation 2 is to optimize the model parameters so as to fit the video frames, using the following loss function (which we refer to as the observation loss):

$$\mathcal{L}_{obs} = \frac{1}{HWT} \sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{t=1}^{T} ||f_\theta(x, y, t) - \mathcal{V}_{xyt}||^2 \tag{4}$$

However, we found that optimizing the INR to only minimize this observation loss leads to overfitting the observation with high temporal frequencies: the intra-frame signal, which we aim to correctly recover, shows important deviations from the observed frames, as illustrated in Figure 2. This observation has led us to consider fitting not only the signal itself, but also to constrain its derivatives. In particular, we regularize the model so as to respect the optical flow constraint. The optical flow constraint equation states that for an infinitesimal lapse of time δt, the brightness of physical points perceived by a camera at arbitrary coordinates (x, y, t) should remain constant. In other words, given the displacement $(\delta x, \delta y)$ of a physical point in the image coordinate system, the image brightness v should remain constant:

$$v(x, y, t) = v(x + \delta x, y + \delta y, t + \delta t) \tag{5}$$

We introduce the vector notation $\mathbf{x} = (x, y, t)$ for readability. Expressing movement as a ratio of displacement in time, we can write the optical flow F and the above constraint as:

$$\begin{aligned} F(\mathbf{x}) &= (\frac{\delta x}{\delta t}, \frac{\delta y}{\delta t}, 1) \\ v(\mathbf{x}) &= v(\mathbf{x} + F(\mathbf{x})) \end{aligned} \tag{6}$$

The first-order Taylor expansion of Equation 6 yields the following:

$$\begin{aligned} v(\mathbf{x}) &= v(\mathbf{x}) + \frac{\delta v}{\delta \mathbf{x}} \cdot F(\mathbf{x}) \\ \frac{\delta v}{\delta \mathbf{x}} \cdot F(\mathbf{x}) &= 0 \end{aligned} \tag{7}$$

which holds exactly in the limit of an infinitesimal δt. We constrain the SIREN derivatives to obey the constraint of Equation 7. Denoting the derivatives of the SIREN as:

$$D_\theta(x, y, t) = \Big(\frac{\delta f_\theta(x, y, t)}{\delta x}, \frac{\delta f_\theta(x, y, t)}{\delta y}, \frac{\delta f_\theta(x, y, t)}{\delta t}\Big) \tag{8}$$

We can now define the optical flow regularization loss as follows:

$$\mathcal{L}_{of} = \frac{1}{HWT} \sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{t=1}^{T} |D_\theta(x, y, t) \cdot F(x, y, t)| \tag{9}$$

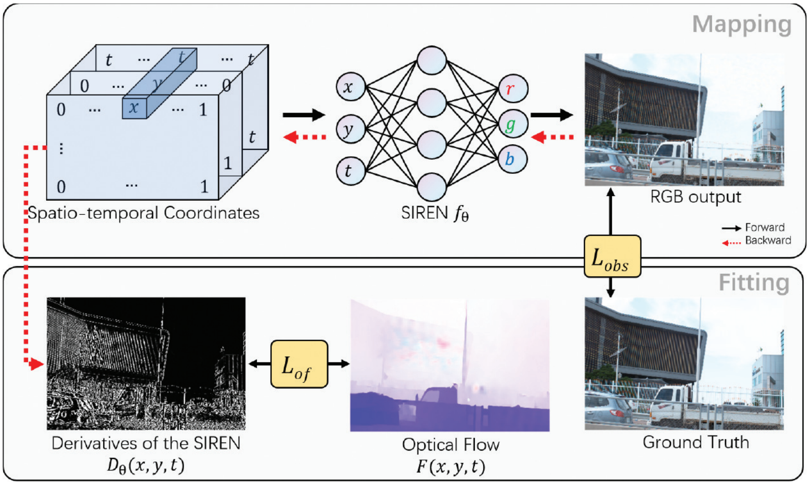


Figure 1: Illustration of our approach. We optimize SIREN to minimize the weighted sum of two losses: The observation loss $\mathcal{L}_{obs}$ measures the fit to the video frames, and the OF loss $\mathcal{L}_{of}$ measures the orthogonality between the SIREN derivatives and the video's optical flow.

The visualizations of $D_\theta(x, y, t)$ and $F(x, y, t)$ are shown in Figure 1. This loss constrains the derivatives of the signal to be orthogonal to the optical flow and can be understood as keeping constant brightness along the optical flow directions. The total loss we use to optimize the INR is a weighted sum of these two terms:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{obs} + \lambda\mathcal{L}_{of} \tag{10}$$

where λ is a hyper-parameter taking values between 0 and 1, whose impact we investigate in the following section. The exactness of the optical flow constraint at the infinitesimal scale plays in our favor: As we regularize the true derivative of the signal representation, we do not assume constant motion on any time interval. We believe this to be the main factor behind our positive results. On the other hand, the optical flow we use was estimated from discrete consecutive frames, and thus does not represent the true infinitesimal motion field but an estimation of finite differences. We discuss potential alternatives in Section 5.

4 Experiments

Following previous works, we use the Adobe [26], X4K [24], Vimeo90K [28], UCF101 [15] and ND Scene [29] datasets as benchmark to compare our method to state-of-the-art models. We use every two frames of each video as observations, and evaluate the ability of SIREN to interpolate on every

Table 1: Quantitative comparison to state-of-the-art VFI on Standard benchmarks. Results are formatted as PSNR/SSIM.

(a) SOTA video frame Interpolation CNN models comparison

	Adobe-240FPS	X4K	Vimeo90K	UCF101*	UCF101**
Super-SloMo [8]	27.77/0.886	27.38/0.852	32.51/0.924	29.01/0.884	-
RRIN [12]	32.37/0.962	30.70/0.927	34.85/0.961	33.17/0.938	-
BMBC [19]	27.83/0.917	27.42/0.858	31.60/0.945	29.74/0.892	-
AdaCof [11]	35.50/0.968	34.61/0.921	37.47/0.966	35.80/0.939	-
ABME [20]	35.28/0.966	34.30/0.919	**39.11**/**0.976**	**36.03**/**0.940**	-
FILM [21]	35.97/0.971	**35.14**/0.939	38.57/0.973	35.72/0.937	-
SIREN [25]	28.90/0.931	25.69/0.849	26.37/0.751	23.68/0.640	25.07/0.697
Ours	**36.52**/**0.977**	35.06/ **0.944**	30.59/0.871	34.57/0.935	**36.24**/**0.953**

(b) Quantitative Comparison of INR Model Interpolation

	ND Scene
V-NF [17]	23.30/0.726
NSFF [13]	28.03/0.925
CURE [23]	**36.91**/**0.984**
SIREN [25]	15.48/0.215
Ours	29.22/0.921

other (intermediate) frame. For the Adobe dataset, we evaluate our method on the test split of eight videos proposed in previous works [8]. We run all additional experiments on the 720p240fps1.mov video of the Adobe dataset (illustrated in Figure 2). Due to the time-consuming operation of optimizing SIREN representations, we optimize and evaluate all models on a 240×360 pixel resolution, and we restrict the Adobe dataset videos to their first 40 frames. For the Vimeo90K and UCF101 datasets, we selected 10 test videos from each dataset as experimental data, and used the original resolutions of the videos.

Unless specified otherwise, we use the following default parameters: SIREN model with depth 9, width 512 and an ω of 30. We optimize the models with the Adam optimizer using a cosine learning rate with maximum learning rate of 10^{-5} during 5000 epochs. We use $\lambda = 0.12$ for the loss function. We compute the optical flow of videos in original resolution using the GMA [9] OF model.

In Section 4.1, we start by highlighting a trade-off akin to underfitting vs overfitting of the signal high frequencies in vanilla SIREN representations. We show that OF-regularized SIREN outperform the best performing vanilla SIREN, showing that the impact of our proposed OF regularization goes beyond high frequency regularization. In Section 4.2, we quantitatively compare our

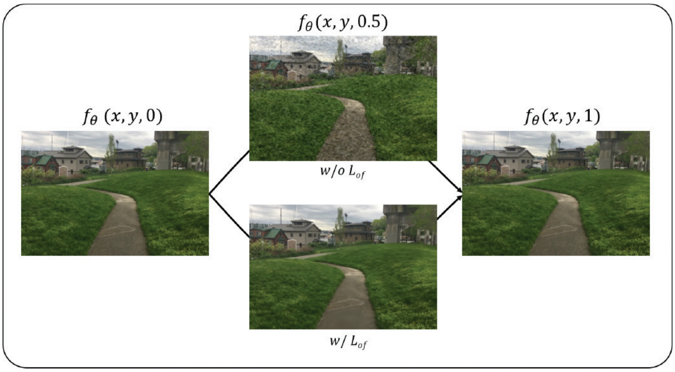


Figure 2: Illustration of INR frame interpolation with and without optical flow regularization. Without regularization (middle top), intermediate frames show unnatural high-frequency variations. Regularizing the INR to satisfy the optical flow constraint equation results in nicely interpolated frames (middle bottom).

method to state of the art models on standard datasets. We show that our method achieves state-of-the-art results on videos with limited motion ranges datasets (Adobe-240FPS, X4K, UCF101), but underperforms recent methods for videos with large motion ranges. Furthermore, through the experimental results, it can be observed that the performance of video frame interpolation significantly degrades when the SIREN model is applied without the optical flow loss. This finding further validates the effectiveness of our proposed method that incorporates optical flow regularization.

We present an ablation study in Section 4.3, providing insights and appropriate settings for the main hyper-parameters, and a qualitative analysis of our results in Section 4.4. Finally, Section 4.5 presents a surprising and counter-intuitive result: we show that our OF loss helps SIREN converge to higher PSNR on the observed frames, opening new potential perspectives for INR optimization and video compressions.

4.1 Optical Flow Constraint and Signal Frequencies

Figure 2 illustrates the fact that the OF constraint smooths out high-frequency noise in the interpolated frames of vanilla SIREN representations.

Healthy skepticism leads us to question whether the impact of the OF constraint is limited to dampening high frequency components of vanilla SIREN representations. To do so, we analyze the representations of vanilla SIREN geared towards different frequency ranges, and compare them to OF-constrained SIREN representations. We constrain the vanilla SIREN frequencies by varying their ω parameter, and report our comparison in Figure 3, with low ω values corresponding to lower frequency ranges.

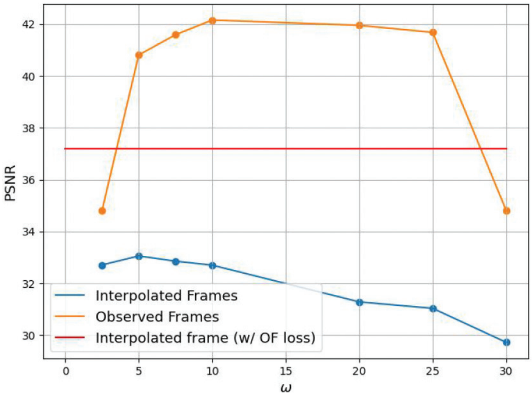


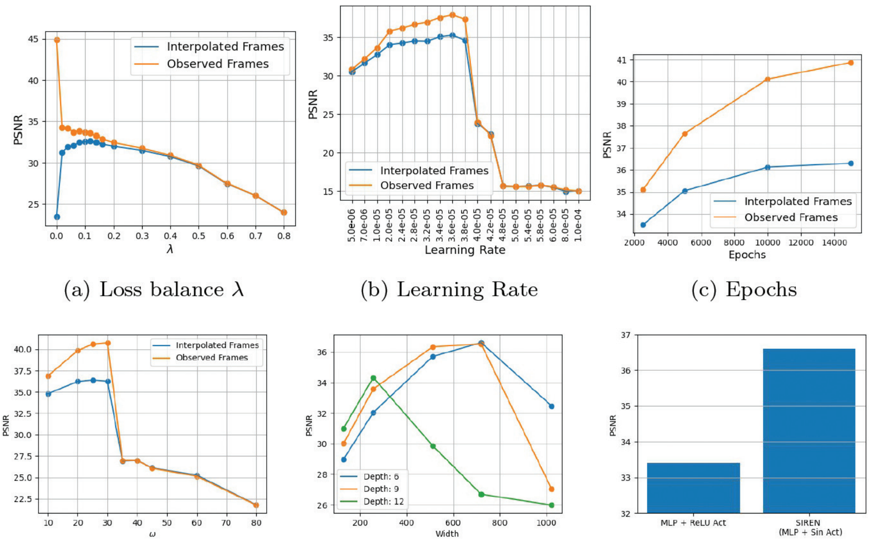
Figure 3: Evolution of the PSNR of observed and interpolated frames with ω without OF loss. Limiting the high frequency fit alone does not reach the same interpolation accuracy as the OF loss.

Constraining the frequency range of vanilla SIREN with ω down to 5 degrades the fit to observed frames but improves interpolation. This suggests that ω behaves similarly to a regularization parameter by controlling a regime of overfitting to the observed frames' high frequencies (for high ω values), versus underfitting (for low ω values). Figure 3 further shows that OF-constrained SIREN achieve far higher interpolation PSNR than the best performing vanilla SIREN, confirming that the OF constraint impact goes beyond dampening of the high frequency noise. Note that in this figure, we kept ω constant for the OF-constrained SIREN to better illustrate our point. The red line represents the results for the best performing ω value. The impact of the ω parameter on OF-constrained SIREN is illustrated separately in Figure 4b.

4.2 State of the Art Models

Table 1 quantitatively compares the results of our method with those of state-of-the-art VFI models on different datasets. Despite its simplicity, and without any training data, our method outperforms most existing models on the Adobe-240FPS and X4K datasets (Table 1). However, as illustrated in Figure 7, it falls short of state-of-the-art methods on the more complex ND Scene benchmark due to larger motion ranges. We provide further comparisons in the qualitative analysis presented in Section 4.4. Section 5 discusses possible strategies for bridging the performance gap on large motion datasets.

It is worth mentioning the Vimeo90K and UCF101 datasets. Currently, the mainstream research on video frame interpolation uses the triplet frames provided by the datasets as the research subject. This means that the in-



(a) Loss balance λ (b) Learning Rate (c) Epochs

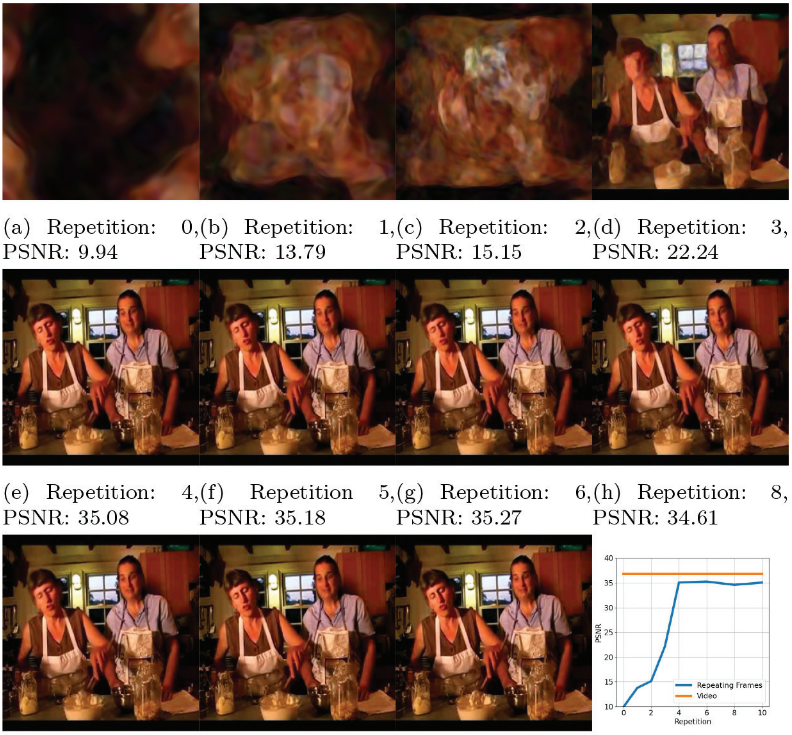
(d) ω (e) Width and Depth (f) Models

Figure 4: Impact of our method's main parameters. Plots from (a) to (d) show both the observed and interpolated frames PSNR while plot (e) and (f) show the interpolated frames PSNR.

put consists of the first and last frames, and the goal is to synthesize the intermediate frame.

However, when using the triplet frames in conjunction with our method, we cannot obtain any meaningful pattern in the interpolated intermediate frames, as shown in Figure 5a. To address this, we attempted to cyclically repeat a triplet video (`v_Mixing_g06_c06.avi`) from the UCF101 dataset, with frame numbering as 0, 1, 2, and repeating as 0, 1, 2, 1, 0, 1, 2, ... and so on. To determine the optimal repetition period, we conducted an ablation experiment, as illustrated in Figure 5i. We found that a repetition period of 6 (resulting in a total of 15 frames) achieved the best PSNR. Therefore, in the results presented in Table 1 for the Vimeo90K and UCF101*, we used a cyclic repetition of triplets with a period of 6 in conjunction with our method to fit the SIREN model. The visualization of the repetition period and the frame interpolation results are shown in Figure 5.

Additionally, besides the triplet frames, UCF101 also provides complete videos from which the triplets are extracted. We were curious to explore whether fitting the model using the video could enhance the frame interpolation results, as more frames would provide a greater representation of continuous motion. To ensure a fair comparison, we used the same number of frames as the triplet repetition video with a period of 6 (i.e., 15 frames) to do the experiment.



(a) Repetition: 0, PSNR: 9.94

(b) Repetition: 1, PSNR: 13.79

(c) Repetition: 2, PSNR: 15.15

(d) Repetition: 3, PSNR: 22.24

(e) Repetition: 4, PSNR: 35.08

(f) Repetition 5, PSNR: 35.18

(g) Repetition: 6, PSNR: 35.27

(h) Repetition: 8, PSNR: 34.61

(i) Repetition: 10, PSNR: 35.07

(j) Video, PSNR: 36.84

(k) Ground Truth

(l) PSNR vs Repetition

Figure 5: Visualization results of the triplet frame video repetition period for the UCF101 dataset.

The visualization of the results is shown in Figure 5j. It is important to note that we only calculated the denoising metrics (PSNR/SSIM) for the frames used for evaluation in the triplet videos, meaning that we only evaluated the denoising performance on a single frame. The experimental results, denoted as UCF101**, are presented in Table 1. The results indicate that fitting the model using the complete videos in the UCF101 dataset leads to better frame interpolation results.

4.3 Ablation Study

Figure 4 summarizes the impact of the main parameters of our method. In (a) we observe a trade-off between the observed and interpolated frames quality in the low λ ranges. The quality of interpolated frames peaks at $\lambda = 0.12$, beyond which point the interpolated frames quality is limited by the quality of the fit to the observed frames, in a similar way to the classical overfitting/underfitting trade-off. However, it should be noted that this trade-off differs widely depending on the SIREN's width. Indeed, as we will show in Section 4.5, the OF constraints actually improves the fit to observed frames for narrow models. In (b) and (c) we observe that both higher learning rates and longer fitting times improve both observed and interpolated frames. The learning rate is limited in amplitude by instabilities of the optimization procedure, while the fitting time is limited by practical time constraints. Large ω (d) also improve the accuracy up to 30, after which instabilities in the optimization see the accuracy drop abruptly. Width and depth (e) show interesting co-dependencies: Increasing width improves interpolation up to a peak after which it degrades. The peak width gets smaller with increasing depth. Models (f) compares the results of frame interpolation with and without the use of SIREN, where both models utilize the OF loss regularization. Replacing SIREN with MLP+ReLU does not result in better frame interpolation performance compared to SIREN. SIREN possesses superior modeling capability for higher-order derivatives of natural signals [25].

Based on these experiments, our final results, as reported in Table 1 were computed with a SIREN model with depth 6, width 720 and $\omega = 25$. We used $\lambda = 0.12$ for the loss, and optimized using Adam with a maximum learning rate of 3.6e-5 during 15k epochs.

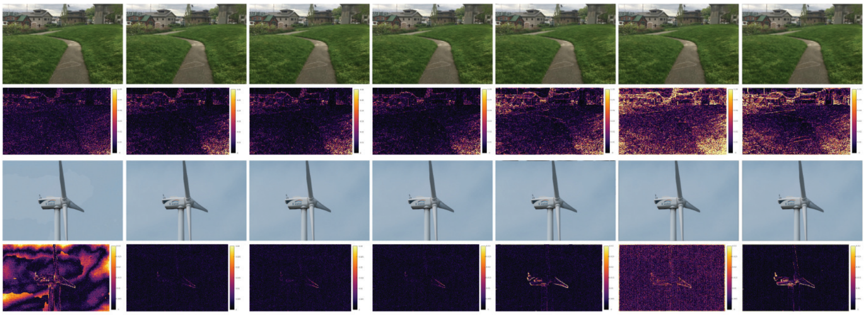
4.4 Qualitative Analysis

Figures 6 and 7 provide a qualitative illustration to the results presented in Section 4.2. The upper frame in Figure 6 shows that our method tends to outperform other methods on videos with limited motion range. In particular, it seems to capture high spatial frequency regions (such as grass and sharp edges of the building) more effectively. In contrast, large motion as illustrated in Figure 7 shows ghosting effects that the OF regularization is not able to address.

The lower part of Figure 6 shows a rare failure case of our method on limited motion ranges: some artificial stain-like patterns appear in the sky background, suggesting additional care may be needed especially in low frequency regions. Despite this rare exception, the overall quality of interpolation on limited motion range videos performs on par with the best existing methods.

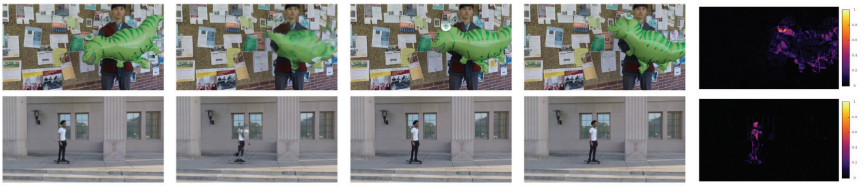
4.5 Video Fitting

Figure 8 shows an unexpected side-effect of the OF regularization observed for narrow networks. As $\mathcal{L}_{obs}$ explicitly maximizes the PSNR of observed frames, we expected the addition of the $\mathcal{L}_{of}$ term to negatively impact the PSNR of



(a) Ours (b) ABME (c) FILM (d) AdaCof (e) BMBC (f) RRIN (g) SloMo

Figure 6: Limited Motion Video Qualitative Analysis (Sampling from Adobe-240FPS, X4K datasets). The interpolated frame results are shown above their residual heat map. The upper frames show a successfully interpolated frames, the lower one shows a rare failure case.



(a) Sampling from ND Scene datasets

(b) $v(x, y, 0)$ (c) $f_\theta(x, y, 0.5)$ (d) $v(x, y, 0.5)$ (e) $v(x, y, 1)$ (f) Residual Error

Figure 7: Large Motion Video Qualitative Analysis (Sampling from ND Scene datasets).

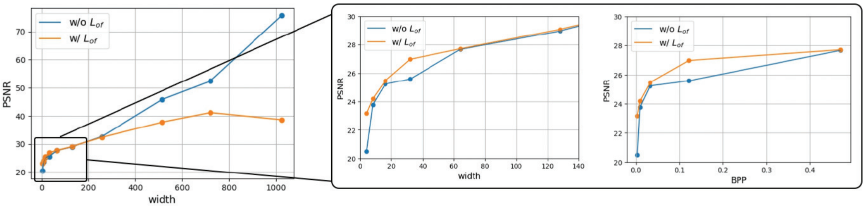


Figure 8: Evolution of the **observed** frames PSNR with depth, with and without OF regularization. Left: Trend from very narrow to very wide models. Right: Zoom on the low width regime with the x axis expressed either in number of neurons or corresponding Bits Per Pixel measure.

observed frames, especially for capacity-limited SIREN which should have to compromise between satisfying both loss terms. It turns out that, for width up to 50, optimizing the SIREN with the additional OF constraint actually improves the fit to observed frames.

Although a complete investigation of this phenomenon is out of the scope of this work, we highlight how this observation may prove interesting for future works: From a practical standpoint, improving the fit of low-capacity INR is the key challenge towards practical INR video compression. It remains to be seen whether this phenomenon can be replicated on more practical architectures (i.e. [3]). From a theoretical standpoint, increasing width has been shown to help optimization by alleviating second order effects [14] and guarantee convergence of gradient descent to global minima [4]. As the understanding of gradient descent dynamics in the high curvature low width setting is currently an elusive question, understanding how the OF constraint helps optimization may provide useful insights into gradient descent dynamics in narrow models.

5 Current Limitations and Future Work

While our method does achieve state-of-the-art interpolation results on Adobe-240FPS, X4K and UCF101 datasets, this work is not meant to deliver a production ready VFI system, which would require the ability to interpolate high resolution and large motion range videos. Instead, we aim to provide insights for future works on both VFI and INR to integrate and build upon. Towards that goal, we discuss below what we see as the three main limitations of our method in its current form, and possible ways to address these limitations.

Slow optimization process. Fitting 20 frames of a video at 240×360 resolution currently takes 15 hours on a 2080Ti GPU using Pytorch. This computation time is an important drawback as it limits our ability to process full resolution video, as well as to explore different hyper parameters and variations of the method within realistic times. We expect advances in INR optimization to be beneficial to this line of research. Given recent successes of INR in signal compression [3, 5, 6, 16, 18, 31], we hopefully expect to see such development in the near future.

Reliance on trained optical flow model. SIREN models allow us to apply the optical flow on the exact derivatives of the signal, bypassing the heuristics of classical methods without relying on machine learning. The optical flow we use, however, is computed by a ML model trained on discrete representations, which raises two problems: it is subject to generalization errors, and is subject to finite difference errors such as occlusions. Bypassing this reliance on ML-based OF using alternative constraints on the exact derivatives of the representation is another interesting way forward.

Inability to interpolate large motion range videos. In its current form, we only apply the optical flow constraint on the observed frames of the video. This has proven sufficient to reach state-of-the art on limited motion ranges, but is not sufficient for large motions. A promising axis of improvement would be to apply additional constraints to the interpolated frames (e.g. for intra-frame time indices $t = 0.5$). Possible regularization methods may include constraints on intra-frame texture, as proposed in recent works [21], or interpolated optical flows, which may prevent the ghosting effects illustrated in Figure 7.

Training a separate model for each video. The current SOTA video frame interpolation CNN models aim to train a model that can be reused for any video, even for unseen scenes, after training on the dataset. Our approach necessitates fitting each video from scratch, with a certain requirement for the number of frames to be fitted, and cannot handle video frame interpolation tasks for unseen scenes. This limitation restricts the practical application range of our method. However, the approach of fitting and interpolating solely with the information available within a single video does provide a novel direction for the field of video frame interpolation. Furthermore, exploring the combination of our proposed method with CNN models could be a meaningful research direction.

6 Conclusion

In this paper, we have shown that SIREN representations of videos can be constrained to satisfy the OF constraint equation in their exact derivatives. We have seen that OF-constrained SIREN reach state of the art VFI on limited motion ranges, without relying on ML based residual flow and interpolation. We have also shown that the OF constraint not only allows SIREN to generate intermediate frames, but can also improve the ability of narrow SIREN to fit observed frames. We have discussed the limitations of our approach in its current form and outlined potentially impactful way forwards for future research.

References

[1] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, “A database and evaluation methodology for optical flow,” *International journal of computer vision*, 92(1), 2011, 1–31.

[2] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, “Performance of optical flow techniques,” *International journal of computer vision*, 12(1), 1994, 43–77.

[3] H. Chen, B. He, H. Wang, Y. Ren, S. N. Lim, and A. Shrivastava, “Nerv: Neural representations for videos,” *Advances in Neural Information Processing Systems*, 34, 2021.

[4] S. S. Du, X. Zhai, B. Poczos, and A. Singh, "Gradient descent provably optimizes over-parameterized neural networks," *arXiv preprint arXiv:1810.02054*, 2018.

[5] E. Dupont, A. Goliński, M. Alizadeh, Y. W. Teh, and A. Doucet, "Coin: Compression with implicit neural representations," *arXiv preprint arXiv:2103.03123*, 2021.

[6] E. Dupont, H. Loya, M. Alizadeh, A. Goliński, Y. W. Teh, and A. Doucet, "COIN++: Data Agnostic Neural Compression," *arXiv preprint arXiv:2201.12904*, 2022.

[7] E. Herbst, S. Seitz, and S. Baker, "Occlusion reasoning for temporal interpolation using optical flow," *Department of Computer Science and Engineering, University of Washington, Tech. Rep. UW-CSE-09-08-01*, 2009.

[8] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super slomo: High quality estimation of multiple intermediate frames for video interpolation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, 9000–8.

[9] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, "Learning to estimate hidden motions with global motion aggregation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 9772–81.

[10] J. Kim, Y. Lee, S. Hong, and J. Ok, "Learning continuous representation of audio for arbitrary scale super resolution," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 3703–7.

[11] H. Lee, T. Kim, T.-y. Chung, D. Pak, Y. Ban, and S. Lee, "Adacof: Adaptive collaboration of flows for video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 5316–25.

[12] H. Li, Y. Yuan, and Q. Wang, "Video frame interpolation via residue refinement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, 2613–7.

[13] Z. Li, S. Niklaus, N. Snavely, and O. Wang, "Neural scene flow fields for space-time view synthesis of dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 6498–508.

[14] C. Liu, L. Zhu, and M. Belkin, "On the linearity of large non-linear models: when and why the tangent kernel is constant," *Advances in Neural Information Processing Systems*, 33, 2020, 15954–64.

[15] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, 4463–71.

[16] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, “Occupancy networks: Learning 3d reconstruction in function space,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 4460–70.

[17] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” in *European conference on computer vision*, Springer, 2020, 405–21.

[18] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, “Deepsdf: Learning continuous signed distance functions for shape representation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 165–74.

[19] J. Park, K. Ko, C. Lee, and C.-S. Kim, “Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation,” in *European Conference on Computer Vision*, Springer, 2020, 109–25.

[20] J. Park, C. Lee, and C.-S. Kim, “Asymmetric bilateral motion estimation for video frame interpolation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 14539–48.

[21] F. Reda, J. Kontkanen, E. Tabellion, D. Sun, C. Pantofaru, and B. Curless, “Film: Frame interpolation for large motion,” in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, Springer, 2022, 250–66.

[22] D. Rho, J. Cho, J. H. Ko, and E. Park, “Neural Residual Flow Fields for Efficient Video Representations,” *arXiv preprint arXiv:2201.04329*, 2022.

[23] W. Shangguan, Y. Sun, W. Gan, and U. S. Kamilov, “Learning Cross-Video Neural Representations for High-Quality Frame Interpolation,” *arXiv preprint arXiv:2203.00137*, 2022.

[24] H. Sim, J. Oh, and M. Kim, “XVFI: Extreme video frame interpolation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 14489–98.

[25] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, “Implicit neural representations with periodic activation functions,” *Advances in Neural Information Processing Systems*, 33, 2020, 7462–73.

[26] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, “Deep video deblurring for hand-held cameras,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 1279–88.

[27] X. Xu, L. Siyao, W. Sun, Q. Yin, and M.-H. Yang, “Quadratic video interpolation,” *Advances in Neural Information Processing Systems*, 32, 2019.

[28] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, “Video enhancement with task-oriented flow,” *International Journal of Computer Vision*, 127, 2019, 1106–25.

[29] J. S. Yoon, K. Kim, O. Gallo, H. S. Park, and J. Kautz, “Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 5336–45.

[30] Y. Zhang, C. Wang, and D. Tao, “Video frame interpolation without temporal priors,” *Advances in Neural Information Processing Systems*, 33, 2020, 13308–18.

[31] Y. Zhang, T. van Rozendaal, J. Brehmer, M. Nagel, and T. Cohen, “Implicit Neural Video Compression,” *arXiv preprint arXiv:2112.11312*, 2021.