Optical Flow Regularization of Implicit Neural Representations for Video Frame Interpolation

Anonymous authors
Paper under double-blind review

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 limited motion ranges using only a target video and its optical flow, without learning the interpolation operator from additional training data. We further show that constraining the INR derivatives not only allows to better interpolate intermediate frames but also improves the ability of narrow networks to fit the observed frames, which suggests potential applications to video compression and INR optimization.

1 Introduction

Many core concepts across the fields of signal processing are defined in terms of continuous functions and their derivatives: surfaces are continuous manifolds in space, motion is a rate of change in space through time, etc. In contrast, modern digital hardware is inherently discrete: digital sensors capture discrete observations of the world regularly sampled in time and space; computers store and process discrete representations of signals. In order to model continuous notions on discrete signal representations, classical methods have used different simplifying assumptions, often taking the form of constant first or second derivatives of the signal between consecutive observations. The lack of generality of any such handcrafted heuristics, combined with the ever improving quantitative results of Machine Learning (ML) methods, have led to the near ubiquitous use of ML in recent signal processing research. These methods leverage large collections of data to infer statistical properties of signals instead of hand-crafted heuristics.

In computer vision, Video Frame Interpolation (VFI) is one task representative of such development. VFI models aim to interpolate intermediate frames between the consecutive frames of a video. To do so, most successful methods rely on the optical flow to guide the interpolation of pixel intensities from the pixel grid of observed frames onto the pixel grid of intermediate frames. Classical methods formulate assumptions such as constant movement or acceleration fields between consecutive frames (Baker et al., 2011; Barron et al., 1994; Herbst et al., 2009). The value of each pixel in the inferred intermediate frame is computed by first shifting the pixel intensities of observed frames along the optical flow directions before interpolating the shifted intensities onto the intermediate frame's pixel grid. Such approaches suffer from the following 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.

These limitations share a common root cause: discretization. Indeed, both the constant brightness assumption, from which is derived the optical flow, and assumptions of constant motion field used by the interpolation process, only truly hold at the infinitesimal scale, for time deltas typically much smaller than those of practically used Frames Per Second (FPS).

ML approaches (Jiang et al., 2018; Li et al., 2020; Park et al., 2021; 2020; Lee et al., 2020) have instead proposed to learn the frame interpolation operator from large video collections, without formulating explicit assumption on the signal. While these approaches have achieved great success in terms of benchmark performance, they are prone to generalization errors caused by domain shifts. Indeed differences between the training set distribution (i.e. VFI benchmark videos) and the target video distribution may hinder the performance of ML models, e.g.; differences stemming from the range of motion, exposure time, FPS and blur (Zhang et al., 2020).

In the mean time, research on implicit representations seeks better discrete representations of continuous signals. In recent years Implicit Neural Representations (INR), i.e. representing signals as Neural Networks (NN) have been shown to offer several competitive advantages over explicit representations, with notable early successes for 3D shape representations (Mildenhall et al., 2020). Of particular interest to us is the work of SIREN (Sitzmann et al., 2020), in which the authors have shown that representing images using Multi Layer Perceptrons (MLP) with sine activation functions allowed for meaningful representations of the signal derivatives. Inspired by this work, we question whether SIREN may be used to guide the interpolation process of VFI by controlling the exact derivatives of the signal instead of the finite differences between consecutive discrete frames, thus avoiding the pitfalls of traditional methods due to discretization. We do so by constraining the derivatives of SIREN representations to satisfy the optical flow constraint equation, i.e., to be orthogonal to the video's optical flow, which we compute using existing state-of-the-art OF models. We find that this method outperforms most existing machine learning-based approaches on small motion range benchmarks, without relying on machine learning for the interpolation operator. In this sense, our method is most similar to classical VIF approaches, except that instead of wrapping the OF on discrete explicit frame representations, we apply the optical flow constraint on the exact gradient of the INR. Our method is thus not subject to any mismatch between training and test data. Furthermore, our approach can sample any number of frame in-between the observed frames due to the continuous nature of the representation. In addition to its application to VFI, we also show that constraining the gradient of the model also improves the ability of narrow MLPs to fit the signal, suggesting potential applications in INR optimization and video compression.

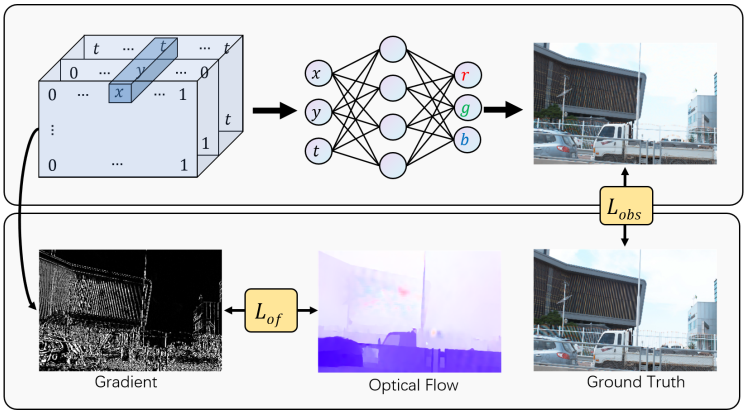


Figure 1: Illustration of our approach. We optimize SIREN to minimize the weighted sum of two losses: The observation loss measures the fit to the video frames, and the OF loss measures the orthogonality between the SIREN derivatives and the video's optical flow.

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 limited motion ranges, 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.

Given 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 to 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

Implicit Neural Representations have met early success in shape representation and 3D rendering (Park et al., 2019; Mescheder et al., 2019; Mildenhall et al., 2020). Since then, a number of works have attempted to apply INR to different signals including audio (Sitzmann et al., 2020; Kim et al., 2022), images (Dupont et al., 2021; 2022), videos (Chen et al., 2021; Shangguan et al., 2022; Rho et al., 2022), medical imaging and climate data (Dupont et al., 2022). In Sitzmann et al. (2020) 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. Combined, 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 works have applied INR to video compression (Zhang et al., 2021; Chen et al., 2021) , with some works Chen et al. (2021) 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 Shangguan et al. (2022), 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 (Baker et al., 2011; Barron et al., 1994; Herbst et al., 2009). Most works 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 Xu et al. (2019), in which the authors propose to take into account acceleration to perform the interpolation, leading to quadratic interpolation. Our work only constrains 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 leverages deep learning for optical flow estimation and interpolation. Super-SloMo (Jiang et al., 2018) 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 (Li et al., 2020) uses residual learning to optimize the performance of Jiang et al. (2018) at the motion estimation bound. AMBE (Park et al., 2021), a current state-of-the-art VFI method, proposes an asymmetric motion estimation method based on Park et al. (2020), which enhances the quality of interpolated frames by loosening the linear motion constraint. Kernel-based approaches such as AdaCof (Lee et al., 2020) 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:

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

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

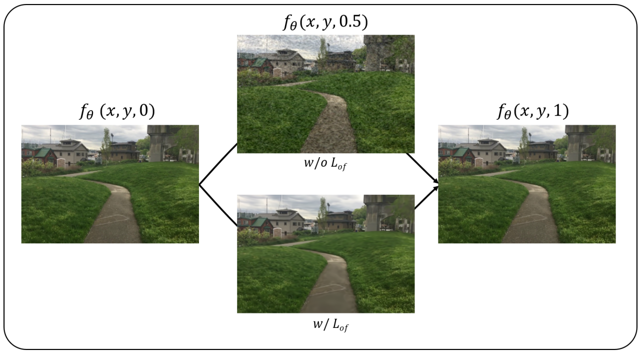


Figure 2: Illustration of INR frame interpolation with and without optical flow regularization. Without regularization (middle top), intermediate frames show unnatural high-frequency variations. Regularizing the INR to satisfy the optical flow constraint equation result in nicely interpolated frames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

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

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 observation 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 $H \times W$ the spatial resolution. We use SIREN as 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 (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 lead us to consider fitting not only the signal itself, but to also 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:

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

First order Taylor expansion of Equation 6 gives 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 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

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

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 (Su et al., 2017), X4K (Sim et al., 2021) and ND Scene (Yoon et al., 2020) 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 other (intermediate) frame. For the Adobe dataset, we evaluate our method on the eight videos test split proposed in previous works (Jiang et al., 2018). 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. 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 uses $\lambda = 0.12$ for the loss function. We compute the optical flow of videos in original resolution using the GMA (Jiang et al., 2021) 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 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, but underperforms recent methods for videos with large motion ranges. 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.

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

(a) Limited motion range

	Adobe-240FPS	X4K
Super-SloMo (Jiang et al., 2018)	27.77 / 0.886	27.38 / 0.852
RRIN (Li et al., 2020)	32.37 / 0.962	30.70 / 0.927
BMBC (Park et al., 2020)	27.83 / 0.917	27.42 / 0.858
AdaCof (Lee et al., 2020)	35.50 / 0.968	34.61 / 0.921
ABME (Park et al., 2021)	35.28 / 0.966	34.30 / 0.919
FILM (Reda et al., 2022)	35.97 / 0.971	**35.14** / 0.939
Ours	**36.52 / 0.977**	35.06/ **0.944**

(b) Large motion range

	ND Scene
V-NF (Mildenhall et al., 2020)	23.30 / 0.726
NSFF (Li et al., 2021)	28.03 / 0.925
CURE (Shangguan et al., 2022)	**36.91 / 0.984**
Ours	29.22 / 0.921

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.

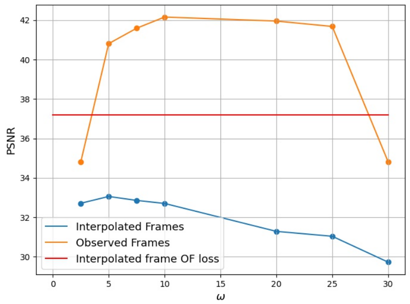


Figure 3: Evolution of the PSNR of observed and interpolated frames with ω without OF loss. Limiting the high frequency fit alone does not reach the same interpolation accuracy as the OF loss.

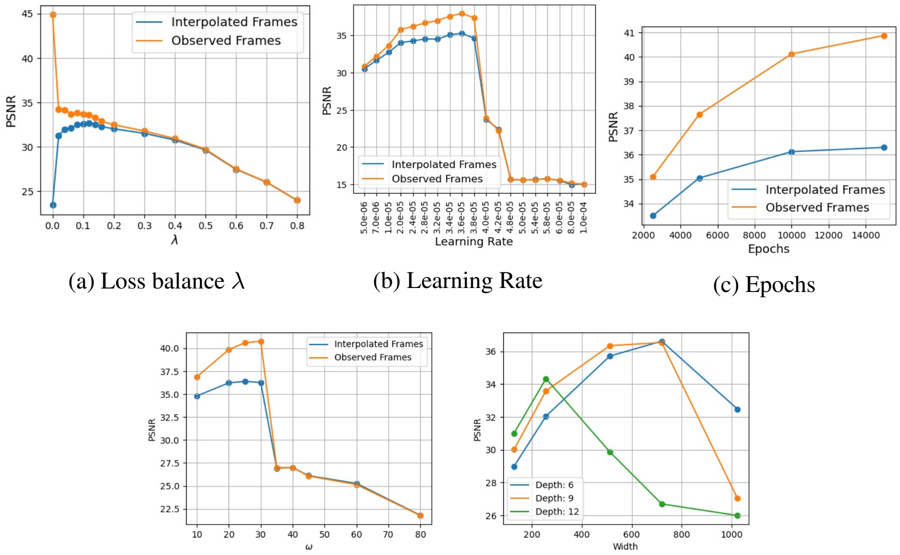
Constraining the frequency range of vanilla SIREN with ω down to 5 degrades the fit to observed frames but improves interpolation. This suggests that ω behaves similarly to a regularization parameter by controlling a regime of overfitting to the observed frames high frequencies (for high ω values), versus 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 we did not vary the ω of the OF constrained SIREN in this figure in order to better illustrate our point, the red line represents results for the best performing ω. The impact of the ω parameter on OF-constrained SIREN is illustrated separately in Figure 4d.

4.2 State of the art models

Table 1 quantitatively compares the results of our method to state-of the art VFI models on different datasets. Despite its simplicity, and without any training data, our method outperforms most existing models on limited motion ranges (Table 1). However, as illustrated in Figure 6, it falls short of state-of-the-art methods on the more complex ND Scene benchmark due to larger motion ranges. We provide further comparison in the qualitative analysis of Section 4.4 and Section 5 discusses possible ways forward to bridging the gap performance on large motion datasets.

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.



(a) Loss balance λ

(b) Learning Rate

(c) Epochs

(d) ω

(e) Width and Depth

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) only shows the interpolated frames PSNR.

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 5 and 6 provide a qualitative illustration to the results presented in Section 4.2. The upper frame in Figure 5 shows that our method tends to outperform other methods on videos with limited motion range. In particular it seems to better catch high spatial frequency regions (grass, sharp edges of the building). In contrast, large motion as illustrated in Figure 6 shows ghosting effects that the OF regularization is not able to address.

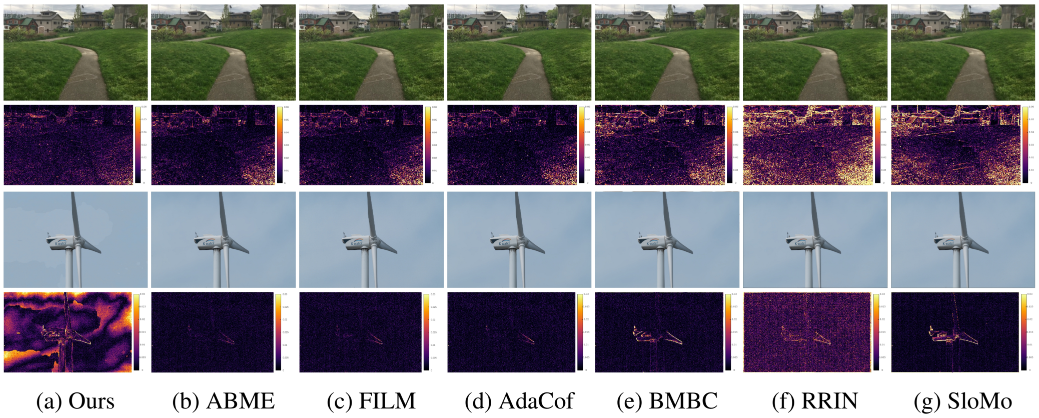


Figure 5: Small Motion Video Qualitative Analysis. The interpolated frame results are shown above their residual heat map. The upper frames show a successfully interpolated frames, the lower one shows a rare failure case.

The lower part of Figure 5 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.

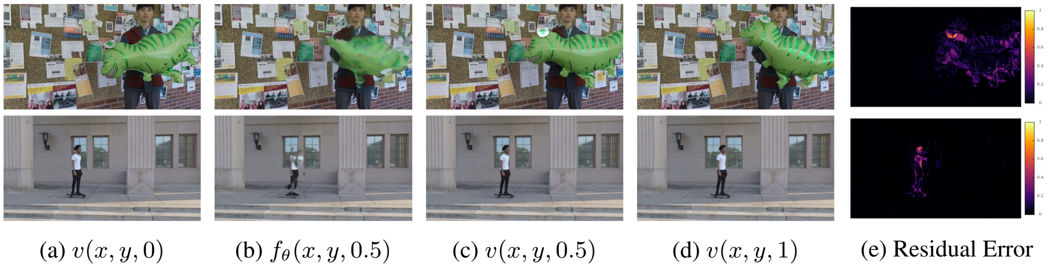


Figure 6: Large Motion Video Qualitative Analysis

4.5 Video fitting

Figure 7 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 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. Chen et al. (2021)). From a theoretical standpoint, increasing width has been shown to help optimization by alleviating second order effects (Liu et al., 2020) and guarantee convergence of gradient descent to global minima (Du et al., 2018). 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.

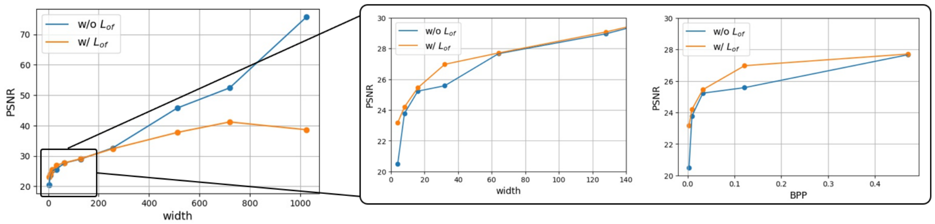


Figure 7: Evolution of the **observed** frames PSNR with depth, with and without OF regularization. Left: Trend from very narrow to very wide models. Right: Zoom on the low width regime with the x axis expressed either in number of neurons or corresponding Bits Per Pixel measure.

5 Current Limitations and Future Work

While our method does reach state of the art interpolation results on limited motion ranges, 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 (Zhang et al., 2021; Dupont et al., 2021; 2022; Park et al., 2019; Mescheder et al., 2019; Chen et al., 2021), 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 (Reda et al., 2022), or interpolated optical flows, which may prevent the ghosting effects illustrated in Figure 6.

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

Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92(1):1–31, 2011.

John L Barron, David J Fleet, and Steven S Beauchemin. Performance of optical flow techniques. *International journal of computer vision*, 12(1):43–77, 1994.

Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser Nam Lim, and Abhinav Shrivastava. Nerv: Neural representations for videos. *Advances in Neural Information Processing Systems*, 34, 2021.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Emilien Dupont, Adam Goliński, Milad Alizadeh, Yee Whye Teh, and Arnaud Doucet. Coin: Compression with implicit neural representations. *arXiv preprint arXiv:2103.03123*, 2021.

Emilien Dupont, Hrushikesh Loya, Milad Alizadeh, Adam Goliński, Yee Whye Teh, and Arnaud Doucet. Coin++: Data agnostic neural compression. *arXiv preprint arXiv:2201.12904*, 2022.

Evan Herbst, Steve Seitz, and Simon 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.

Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan 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*, pp. 9000–9008, 2018.

Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9772–9781, 2021.

Jaechang Kim, Yunjoo Lee, Seunghoon Hong, and Jungseul 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)*, pp. 3703–3707. IEEE, 2022.

Hyeongmin Lee, Taeoh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, and Sangyoun Lee. Adacof: Adaptive collaboration of flows for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5316–5325, 2020.

Haopeng Li, Yuan Yuan, and Qi Wang. Video frame interpolation via residue refinement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2613–2617. IEEE, 2020.

Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver 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*, pp. 6498–6508, 2021.

Chaoyue Liu, Libin Zhu, and Misha Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. *Advances in Neural Information Processing Systems*, 33: 15954–15964, 2020.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4460–4470, 2019.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pp. 405–421. Springer, 2020.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 165–174, 2019.

Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation. In *European Conference on Computer Vision*, pp. 109–125. Springer, 2020.

Junheum Park, Chul Lee, and Chang-Su Kim. Asymmetric bilateral motion estimation for video frame interpolation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14539–14548, 2021.

Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. *arXiv preprint arXiv:2202.04901*, 2022.

Daniel Rho, Junwoo Cho, Jong Hwan Ko, and Eunbyung Park. Neural residual flow fields for efficient video representations. *arXiv preprint arXiv:2201.04329*, 2022.

Wentao Shangguan, Yu Sun, Weijie Gan, and Ulugbek S Kamilov. Learning cross-video neural representations for high-quality frame interpolation. *arXiv preprint arXiv:2203.00137*, 2022.

Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. Xvfi: Extreme video frame interpolation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14489–14498, 2021.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1279–1288, 2017.

Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. *Advances in Neural Information Processing Systems*, 32, 2019.

Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan 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*, pp. 5336–5345, 2020.

Youjian Zhang, Chaoyue Wang, and Dacheng Tao. Video frame interpolation without temporal priors. *Advances in Neural Information Processing Systems*, 33:13308–13318, 2020.

Yunfan Zhang, Ties van Rozendaal, Johann Brehmer, Markus Nagel, and Taco Cohen. Implicit neural video compression. *arXiv preprint arXiv:2112.11312*, 2021.