Supplementary: Non-linear Motion Estimation for Video Frame Interpolation using Space-time Convolutions

Saikat Dutta Arulkumar Subramaniam Anurag Mittal
Indian Institute of Technology Madras
Chennai, India

1. GridNet-3D description

GridNet-3D consists of three parallel streams to capture features with different resolutions and each stream has five convolutional blocks arranged in a sequence as shown in Fig. 1. Each convolutional block is made up of two conv-3D layers with a residual connection. The three parallel streams have channel dimensions of 16, 32 and 64 respectively. The communication between the streams are handled by a set of *downsampling* and *upsampling* blocks. The *downsampling* block consists of spatial max pooling of stride 2 followed by one conv-3D layer, whereas the *upsampling* block consists of one bilinear upsampling layer followed by two conv-3D layers.

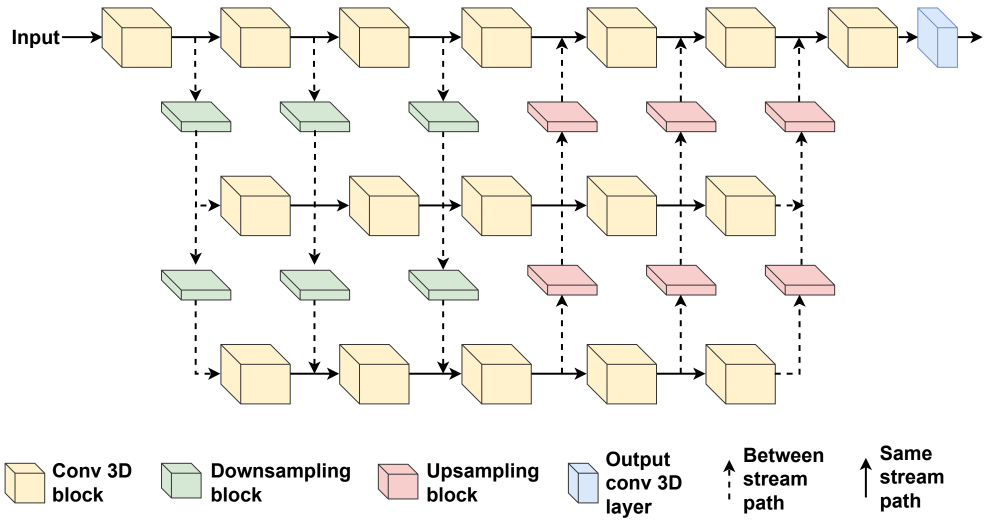


Figure 1. Novel GridNet-3D architecture for efficient multi-scale feature aggregation inspired from [10]. It consists of three parallel streams operating in different feature resolutions and the communication between streams is handled by *downsampling* and *upsampling* blocks.

2. Dataset description

Vimeo Septuplet dataset: Vimeo Septuplet dataset [9] consists of 72,436 frame-septuplets of resolution 256×448. This dataset is divided into a training subset of 64,612 septuplets and a test subset of 7,824 septuplets. We use 1st, 3rd, 5th and 7th frame from the septuplets as input frames and predict the 4th frame as interpolation ground truth. We use training subset of this dataset for training and evaluate the model on other datasets without fine-tuning.

DAVIS dataset: DAVIS-2017 TrainVal dataset [6] contains 90 video clips with diverse scenes and complex motions. We utilize its 480p counterpart for evaluation purposes. We extract 2,849 quintuplets from the provided video sequences.

HD dataset: Bao et al. [1] collected 11 HD videos consisting of four 544p, three 720p and four 1080p videos. We extract 456 quintuplets from these videos and discard 8 quintuplets with blank frames and scene changes. Finally, we use 448 quintuplets for evaluation.

GoPro dataset: GoPro dataset proposed by Nah et al. [5] contains 33 720p videos captured at 720 FPS. We extract 1,500 sets of 25 images from the test split consisting of 11 videos. We use 1st, 9th, 17th and 25th frames as input frames and 13th frame is used as the interpolation target.

3. Experiments on model configurations

In this section, we perform comparative studies among the different choices available for NME (UNet-2D [3], UNet-3D [4], GridNet-3D) and MR (UNet-2D [7], GridNet-2D [2]) modules to determine the best performing configuration.

Choice of NME module: We experiment with three different architectures for NME module: 1) UNet-2D [3], 2) UNet-3D [4], and 3) novel GridNet-3D proposed in this paper. We illustrate the quantitative performance with different NME modules in Table 1 along with number of parameters and runtimes. We observe that 3D-CNN version of NME modules perform superior to UNet-2D in general. Further, GridNet-3D performs better than UNet-3D in DAVIS, HD and GoPro datasets while having less parameters and runtime.

Table 1. Effect of different CNN architectures used in NME module. Best and second best scores are colored in red and blue respectively.

CNN used	Vimeo Septuplet		DAVIS		HD		GoPro		Params	Runtime
in NME	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	(M)	(s)
UNet-2D	34.76	0.9537	27.34	0.8254	31.21	0.8971	28.90	0.8793	38.30	0.18
UNet-3D	34.96	0.9545	27.46	0.8278	31.31	0.8976	29.01	0.8826	60.55	0.37
GridNet-3D	34.99	0.9544	27.53	0.8281	31.49	0.9000	29.08	0.8826	20.92	0.32

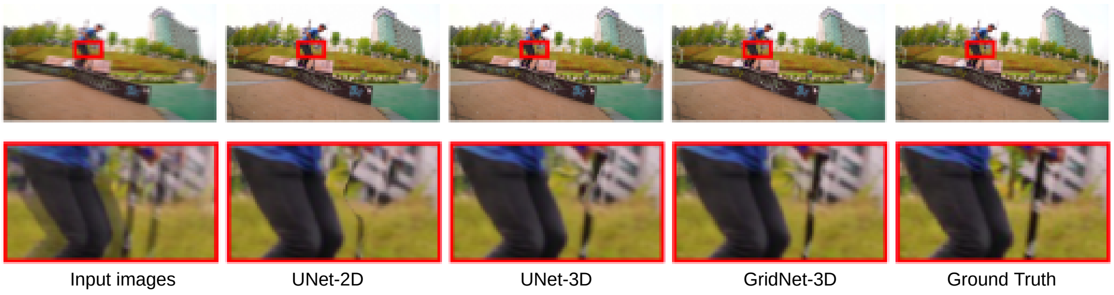


Figure 2. Qualitative comparison between different CNN architectures used in NME module.

Choice of MR modules: We experiment with two types of motion refinement modules: UNet-2D [7] and GridNet-2D [2]. We use a standard encoder-decoder architecture with skip connections for UNet-2D. In GridNet-2D, encoder and decoder blocks are laid out in a grid-like fashion to carry through multi-scale feature maps till the final layer. Quantitative comparison in Table 2 shows that using GridNet-2D as MR module performs significantly better than UNet-2D. Qualitative comparison in Figure 3 illustrates that GridNet-2D reduces the smudge effect in interpolated frame compared to UNet-2D. From Table 2, we can also infer that using GridNet-2D as MR module reduces total number of parameters of the model while the runtime remains constant.

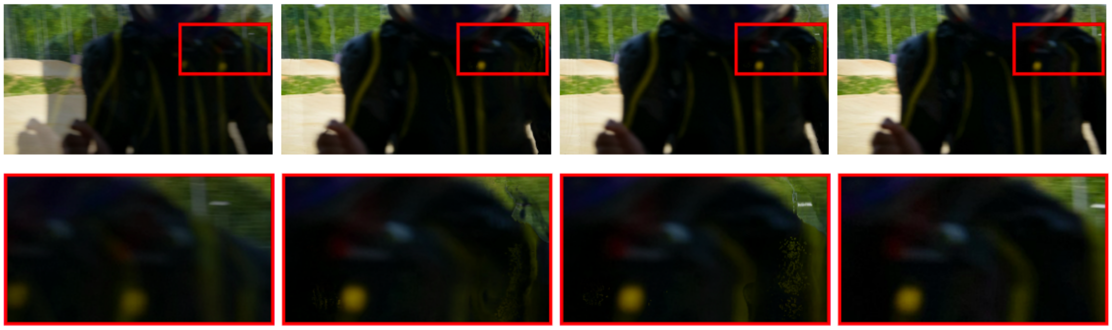
Based on these experiments, we use GridNet-3D in NME module and GridNet-2D as MR module in state-of-the-art comparisons and in ablation studies unless specified otherwise.

Table 2. Quantitative comparison between UNet and GridNet as MR module.

Motion Refinement	Vimeo Septuplet		DAVIS		HD		GoPro		Params	Runtime
module	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	(M)	(s)
UNet-2D	34.70	0.9532	27.32	0.8260	31.02	0.8944	28.81	0.8798	78.11	**0.37**
GridNet-2D	**34.96**	**0.9475**	**27.46**	**0.8278**	**31.31**	**0.8976**	**29.01**	**0.8826**	**60.55**	**0.37**

4. Ablation studies

Choice of input features (RGB vs. Flow+Occlusion): To demonstrate the importance of flow and occlusion maps, we perform an experiment where we use RGB frames as input to the 3D CNN. Quantitative comparison between these two approaches are shown in Table 3 with number of parameters and runtimes. Both experiments in Table 3 use UNet-2D as MR



Input Images

UNet-2D as MR module

GridNet-2D as MR module

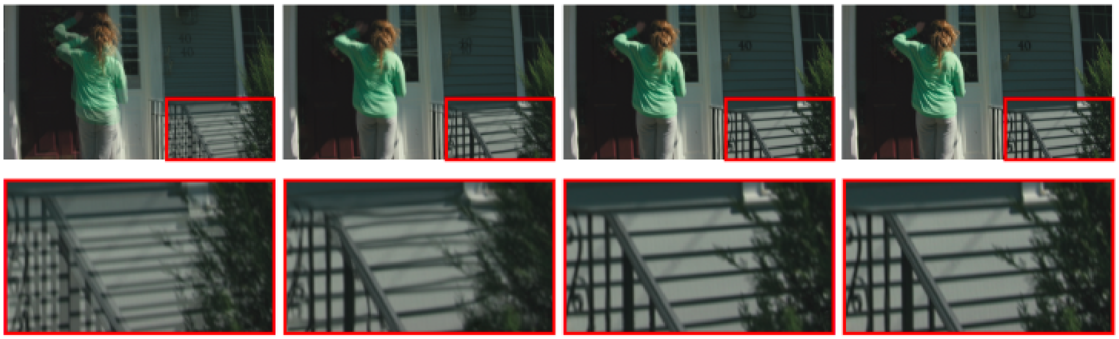
Ground Truth

Figure 3. Qualitative comparison between different MR modules.

module. We observe that Flow+Occlusion maps as input performs better than RGB frames. Qualitative comparison in Figure 4 shows that interpolated results are more accurate when Flow+Occlusion maps are used compared to RGB. Note that, our model with RGB input already performs better than FLAVR [4] (refer to Table 1 of main paper). This signifies that frame generation by hallucinating pixels from scratch [4] is hard for neural networks to achieve than frame generation by warping neighborhood frames.

Table 3. Effect of different input features to 3D CNN.

Input	Vimeo Septuplet		DAVIS		HD		GoPro		Params	Runtime
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	(M)	(s)
RGB	34.12	0.9474	26.34	0.7883	30.80	0.8854	28.34	0.8642	**61.89**	**0.23**
Flow + Occlusion	**34.70**	**0.9532**	**27.32**	**0.8260**	**31.02**	**0.8944**	**28.81**	**0.8798**	78.11	0.37



Input frames

RGB as input

Flow + Occlusion as input

Ground Truth

Figure 4. Qualitative comparison between RGB and Flow+Occlusion as input to 3D CNN.

Importance of BFE, MR and BME modules: To understand the importance of BFE, MR and BME modules, we re-purpose the NME module to directly predict non-linear backward flows $F_{t\rightarrow 0}$, $F_{t\rightarrow 1}$ and blending mask M. In this experiment, we use RGB frames as input to NME module. The quantitative comparison in Table 4 illustrates that the direct estimation of backward flows, mask (without BFE, MR and BME) performs sub-par to estimating them with BFE, MR (UNet-2D) and BME modules. Added to this, qualitative comparisons in Figure 5 shows that the direct estimation of backward flows may lead to ghosting artifacts due to inaccurate flow estimation near motion boundaries.

Table 4. Quantitative significance of BFE, MR and BME modules.

	Vimeo Septuplet		DAVIS		HD		GoPro		Params	Runtime
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	(M)	(s)
without BFE, MR and BME	33.91	0.9443	26.05	0.7686	30.72	0.8811	28.12	0.8583	**42.07**	**0.20**
with BFE, MR and BME	**34.12**	**0.9474**	**26.34**	**0.7883**	**30.80**	**0.8854**	**28.34**	**0.8642**	61.89	0.23

Importance of using four frames: In order to show the effectiveness of using four frames in our network, we report results of our model using only two frames (I_0, I_1) as input. Since our model expects four frames as input, we use frame

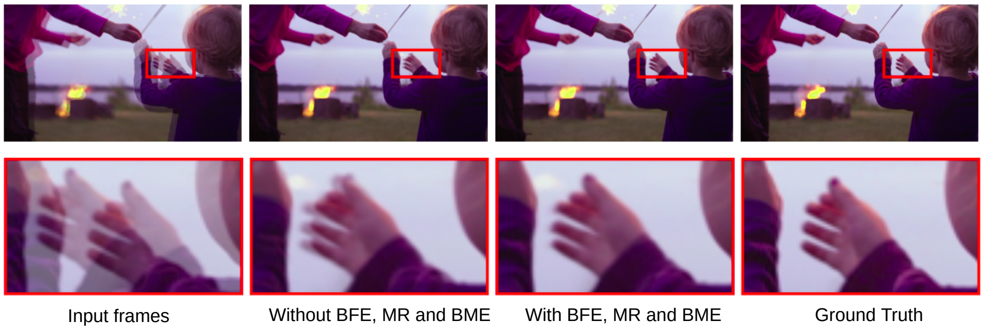


Figure 5. Qualitative comparison between intermediate flowmap and blending mask estimation with and without BFE, MR and BME modules.

repetition (I_0, I_0, I_1, I_1) in this experiment. Quantitative comparison in Table-5 shows that our model indeed benefits from using four frames as input.

Table 5. Quantitative comparison between using two frames and four frames

No. of input frames	Vimeo Septuplet		DAVIS		HD		GoPro	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
2	33.61	0.9438	26.04	0.7737	30.97	0.8901	27.27	0.8347
4	**34.99**	**0.9544**	**27.53**	**0.8281**	**31.49**	**0.9000**	**29.08**	**0.8826**

5. Multi-frame interpolation

We have tested 4x interpolation results (generating 3 intermediate frames) in a recursive way on GoPro test set and compared it with QVI [8]. Quantitative results are shown in Table-6. We can see that our model can perform better than QVI [8] on multi-frame interpolation case too.

Table 6. Multi-frame interpolation results on GoPro dataset.

Method	PSNR	SSIM
QVI	29.36	0.8964
Ours	**29.86**	**0.9021**

6. Parameter and runtime analysis of different components

In Table-7, we have reported number of parameters and average runtime of different components of our network.

Table 7. Component-wise parameter and runtime analysis

Component name	Specification	Params (M)	Runtime (s)
Flow and occlusion estimator	-	16.19	0.12
3D CNN	UNet-3D	42.06	0.20
3D CNN	GridNet-3D	2.44	0.15
BFE	-	0	0.02
MR	UNet	19.81	0.016
MR	GridNet	2.25	0.021
BME	-	0.04	0.002
Frame Synthesis	-	0	0.002

References

[1] Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[2] Damien Fourure, Rémi Emonet, Élisa Fromont, Damien Muselet, Alain Trémeau, and Christian Wolf. Residual conv-deconv grid network for semantic segmentation. *arXiv preprint arXiv:1707.07958*, 2017. 2

[3] 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*, pages 9000–9008, 2018. 2

[4] Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, and Du Tran. Flavr: Flow-agnostic video representations for fast frame interpolation. *arXiv preprint arXiv:2012.08512*, 2020. 2, 3

[5] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 2

[6] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[8] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *NeurIPS*, 2019. 4

[9] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 1

[10] YUE Yuanchen, CAI Yunfei, and WANG Dongsheng. Gridnet-3d: A novel real-time 3d object detection algorithm based on point cloud. *Chinese Journal of Electronics*, 30(5):931–939, 2021. 1