MagFormer: Hybrid Video Motion Magnification Transformer from Eulerian and Lagrangian Perspectives

Sicheng Gao[1,*], Yutang Feng[1,*], Linlin Yang[2], Xuhui Liu[1], Zichen Zhu[4], David Doermann[5], Baochang Zhang[1,3,†]

[1]Beihang University, [2]University of Bonn, [3]Zhongguancun Laboratory, [4]Harbin Institute of Technology, [5]University at Buffalo

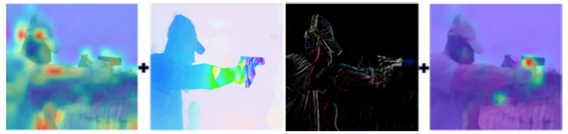
BMVC 2022

Motivation

Video motion magnification can be mainly divided into two categories

· **Eulerian Lagrangian approaches :** Amplify the motions of moving objects using optical flow (e.g. Liu *et al.*[1])

· **Eulerian approaches :** Measure and amplify the variations over time based on the pixel-wise change with fixed spatial locations (e.g. EVM [2], Phase-based [3], VAM [4], Oh et al. [5])



(a) Lagrangian Branch (b) Eulerian Branch

(a) A global attention map and local optical flow in our Lagrangian branch by using quantifying attention method [6].
(b) Global motion flow and a local activation map in our Eulerian branch by using the CAM method [7].

Proposed Method

1. Motion guided attention module

We use the optical flow O and the current input frame I as input, and provide motion magnification attention A in each Transformer and CNN block based on motion- guided attention module.

$$A = (\alpha - 1)\,\mathrm{Sigmoid}(h(I, O)) + \mathbb{1},$$

2. Two-branch module

Lagrangian Branch :

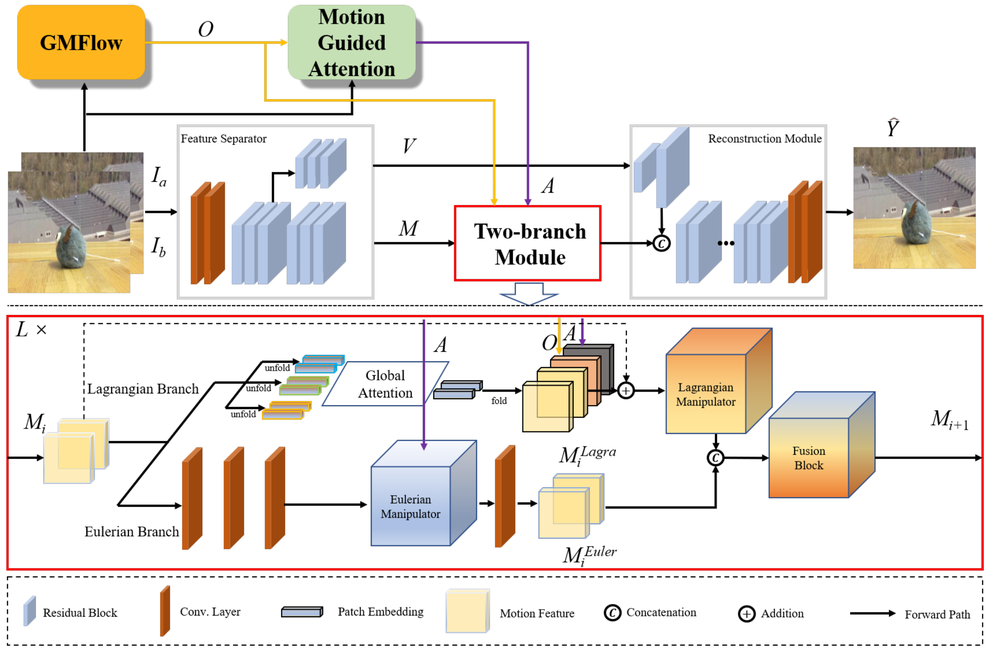
$$M_{i+1}^{Lagra} = \mathrm{Res}[\Delta(\mathrm{LN}(\bar{M}_i), A \odot O)].$$

Here, Res is a residual block to adapt the magnification process and maintain the quality of the magnified frame, ⊙ denotes Hadamard product, LN is the LayerNorm layer, σ is a fusion module.

Eulerian Branch :

$$M_{i+1}^{Euler} = \mathrm{Conv}[M_b + \mathrm{Res}(\mathrm{Conv}(G(M_i) \odot A))],$$

where Conv is a convolutional layer and G(·) means 3 convolutional layers. Also, M_b means the next frame and $M_1 = M_b - M_a$.



Overview of our MagFormer. It contains an optical flow extraction (GMFlow), a motion-guided attention module, a feature separator, a two-branch module and a reconstruction module.

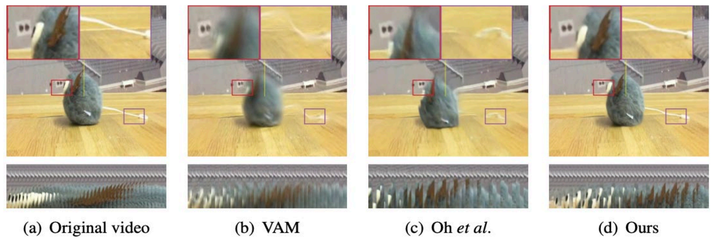
Experiments and Results

Average PSNR and SSIM of all testing videos, using different motion magnification methods with different magnification factors. The presentation format is PSNR / SSIM. The best results are in bold.

PSNR/SSIM	Phase-based	Oh *et al.* (Static)	Oh *et al.* (Dynamic)	Ours
$\alpha = 10$	22.80/0.7777	21.86/0.7446	**26.95/0.8658**	26.46/0.8452
$\alpha = 20$	21.78/0.7235	21.14/0.7106	24.40/0.8217	**25.73/0.8369**
$\alpha = 40$	20.82/0.6776	21.01/0.7005	23.25/0.7968	**25.38/0.8306**

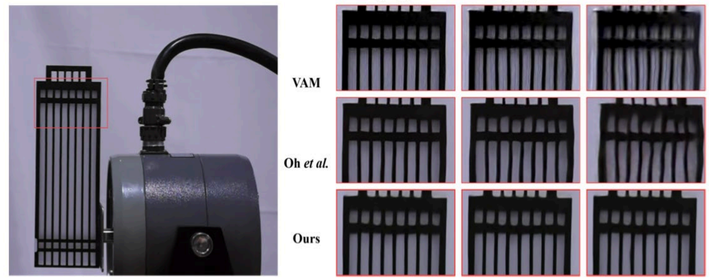
Average PSNR and SSIM of different motion magnification methods of six videos with α = 40. The presentation format is PSNR / SSIM. The best results are in bold.

PSNR/SSIM	Phase-based	Oh *et al.* (Static)	Oh *et al.* (Dynamic)	Ours
cattoy	23.75/0.6808	22.68/0.6836	23.39/0.7099	**29.20/0.8908**
drone	18.57/0.5481	17.08/0.4984	19.51/0.6000	**25.92/0.8156**
bottle	20.46/0.8246	20.26/0.8489	20.27/0.8767	**23.68/0.9088**
eye	20.1/0.8262	25.14/0.8766	**27.46/0.9023**	23.57/0.7832
plants	19.99/0.5432	19.02/0.5752	24.44/**0.8925**	**24.63**/0.7543
drum	22.06/0.6429	21.86/0.7205	24.47/0.7996	**25.30/0.8311**



(a) Original video (b) VAM (c) Oh *et al.* (d) Ours

Cropped frame of the cat toy video when magnification factor is 10. The toy is moving from left to right while vibrating. The top row shows the detail of two sub-regions of the image. The bottle row shows a single column of pixels in the yellow line of the cropped image of the corresponding frames.



Comparison with VAM [4] and Oh et al. [5] on the exciter videos with different α.

Conclusion

1. We propose **a novel unified framework**, MagFormer, for video motion magnification.
2. We introduce **a motion-guided attention module** to highlight the motion areas and reduce the annoying video artifacts
3. We introduce **a hybrid two-branch module** with a Transformer branch from Eulerian perspective and a CNN branch from Lagrangian perspective.
4. We introduce **a new vibration dataset and a corresponding metric** to evaluate video motion magnification quantitatively via amplitude and frequency.

References

[1] C. Liu, A. Torralba, W. Freeman, F Durand, and E. Adelson. Motion magnification. TOG, 2005.
[2] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. TOG, 2012.
[3] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. TOG, 2013.
[4] Yichao Zhang, Silvia L Pintea, and Jan C Van Gemert. Video acceleration magnification. In CVPR, 2017.
[5] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Fr'edo Durand, William T Freeman, and Wojciech Matusik. Learning-based video motion magnification. In ECCV, 2018.
[6] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In ACL, 2020.
[7] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In CVPR, 2016.