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

Sicheng Gao[1*]
scgao@buaa.edu.cn
Yutang Feng[2*]
yutangfeng@buaa.edu.cn
Linlin Yang[3]
yangl@informatik.uni-bonn.de
Xuhui Liu[1]
xhliu@buaa.edu.cn
Zichen Zhu[5]
zzchit@stu.hit.edu.cn
David Doermann[6]
doermann@buffalo.edu
Baochang Zhang[1,4†]
bczhang@buaa.edu.cn

[1] Institute of Artificial Intelligence, School of Automation Science and Electrical Engineering, Beihang University, Beijing, China
[2] Sino-French Engineer School, Beihang University, Beijing, China
[3] University of Bonn, Germany
[4] Zhongguancun Laboratory, Beijing, China
[5] School of Mechatronics Engineering, Harbin Institute of Technology, Harbin, China
[6] University at Buffalo, USA

Abstract

Video motion magnification methods attract much attention for their strong capability of capturing informative subtle signals from diverse engineering scenes. There are two main types of methods in this field, Eulerian and Lagrangian motion magnification, which have different advantages and perspectives. However, the combination of both remains largely unexplored. In this paper, we develop an end-to-end video motion magnification network, MagFormer, with a well-designed two-branch magnification module, which includes a convolutional neural network (CNN) for the Eulerian motion magnification branch and Transformer for the Lagrangian optical flow magnification branch. Our MagFormer can inherit the advantages of two perspectives, by leveraging both Eulerian global motion features from the camera-centered perspective and trajectories of the object-centered from the Lagrangian object perspective in a unified parallel framework. To validate the effectiveness of our method, we collect a new vibration dataset to measure video motion magnification methods via amplitude and frequency. More experiments are conducted on fixed-background subtle motion videos, constantly moving object videos and quantitative vibration videos. Experimental results show that our method achieves a favorable improvement compared to state-of-the-art methods. Codes will be released at https://github.com/Ree1s/MagFormer.

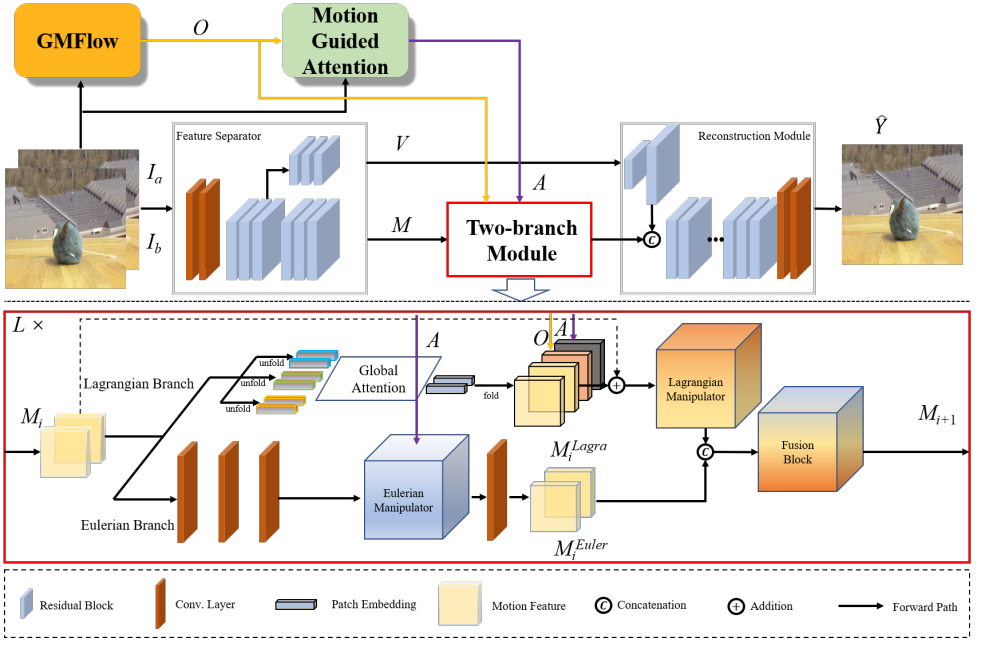


Figure 1: Overview of our MagFormer. It contains an optical flow extraction (GMFlow), a motion-guided attention module, a feature separator, a two-branch module and a reconstruction module.

1 Introduction

Video motion magnification aims to detect and visualize subtle motion signals that conceal valuable information and has drawn much attention in high-precision applications, such as surgical video analysis [9] and vibration of building [5]. Analogous to hydrodynamics, the study of video motion magnification is founded on Eulerian and Lagrangian perspectives. Eulerian approaches measure and amplify the variations over time based on the pixel-wise change with fixed spatial locations. In contrast, Lagrangian approaches discern small motions by adopting explicit tracking of pixels and extracting the optical flow.

Eulerian approaches benefit from their spatio-temporal analysis and high efficiency by fixing the perspective to compute the motion flux [21, 23]. However, they require a delicate design of signal frequency knowledge and decomposition filters. Lagrangian approaches focus on the motion trajectory of the pixels of interest in the video and can effectively magnify the range of motion [6, 7, 13]. But the pixel-wise tracking process of Lagrangian methods consumes expensive computational resources and fails to consider global information of the entire images [2]. As Eulerian approaches and Lagrangian approaches are complementary by nature, hybrid methods with both elements are appealing to inherit the advantages of both perspectives. In this case, we introduce a MagFormer network with a two-branch

*S. Gao and Y. Feng contributed equally to this work.
†B. Zhang is the corresponding author.

module of both CNN and Transformer for motion video magnification. The module exploits CNN for the Eulerian motion magnification branch and Transformer for the Lagrangian optical flow magnification branch, to leverage both global motion features from the Eulerian camera-centered perspective and trajectories of the object-centered from the Lagrangian object perspective.

As shown in the upper part of Fig. 1, our MagFormer takes account of the motion magnification attention, texture features and motion features to make the system effective and efficient. Specifically, a feature separator is proposed to extract texture features and motion features from consecutive video frames. Also, the optical flow extractor (GMFlow) and the motion-guided attention module are introduced to calculate motion magnification attention, which shows an excellent improvement of the magnification effect of the moving object. Then, as shown in the bottom part of Fig. 1, we implement a two-branch module to magnify motion features with motion magnification attention in a layer-by-layer manner. It can highlight local magnified optical flow motion features of moving objects and global Eulerian motion features of the background. In order to interact motion features of two branches, we fuse the Eulerian motion branch and the optical flow branch with fusion blocks and achieve better magnification results. Finally, a reconstruction block is introduced to generate the magnified output from texture features and magnified motion features. The main contributions are summarized as follows:

- We implement a two-branch module, including Eulerian motion magnification CNN branch and Lagrangian motion magnification Transformer branch. It interacts with global motion feature flow and local precise optical flow of the moving object and magnifies both in a layer-by-layer manner.
- We introduce an end-to-end video motion magnification framework, called MagFormer, which includes the optical flow extractor, the motion-guided attention module, the feature separator and the reconstruction module. Especially, it integrates texture features and magnified motion features into a unified parallel framework and leads to an effective and efficient network.
- For video motion magnification, we introduce a new vibration dataset collected by a modal exciter and a corresponding metric to measure motion magnification via amplitude and frequency.
- Experiments on previous real-world videos and our newly proposed vibration evaluation dataset demonstrate that our model outperforms prior arts for the quality of output videos and quantitative physical information.

2 Related Work

Early Eulerian approaches [21, 22, 23] simply decompose the input images into different pyramid levels and magnify the motion by choosing the proper filter. For example, to extract the subtle motion representation in the input frames, Wu *et al*. [23] propose the first-order Taylor expansion, while works like [21, 22] use the complex steerable pyramid. To further enhance the motion magnification quality, the work [19] proposes a bilateral video magnification filter (BVMF) which runs two kernels in the temporal and intensity domains. However, those works heavily rely on hand-crafted filters and may simply suppress all the quick

large motion. Recently, beyond hand-crafted filters, learning-based Eulerian [16] proposes to learn the proper parameters from a synthetic dataset and has less edge artifact. Nevertheless, this method is designed in Eulerian perspective, which will magnify all the motions and the noise. When the magnification factor is large, the noise will disturb the valid information seriously. Unlike Eulerian approaches analyzing motion based on fixed positions, lagrangian's methods [13] can explicitly amplify the motions of moving objects using optical flow. However, used naively, the calculation of optical flow in a traditional way will occupy a huge amount of memory, which is unacceptable when the sizes of input images are large. Although the work [4] also aims to combine Eulerian with Lagrangian, it is not end-to-end and limited with poor generalization and a low inference speed.

Learning-based approaches [3, 11, 15, 17, 18] and their network architectures have achieved remarkable progress due to the revolution of deep learning. The work [16] is the only learning-based approach work that adopting CNN [11, 18, 24] for motion magnification. Considering the motion magnification requires both local information to capture the object's motion and global information to retain background motion, we firstly introduce a two-branch module that integrates both CNN with Transformer [3, 20] to extract local and global information for Eulerian branch and Lagrangian branch, respectively. Also, we propose to use a motion-guided mask to select the motion of interest in videos, which can greatly reduce the annoying video artifacts and wrongly magnified motions suffered by traditional Eulerian approaches.

3 Method

3.1 MagFormer

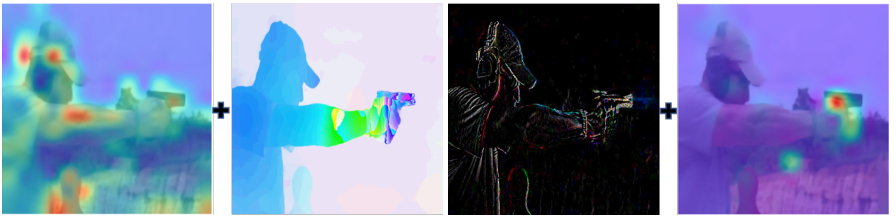
We present an overview of our MagFormer in the upper part of Fig. 1. Our proposed network architecture consists of a pre-trained flow-estimation network (*i.e.*, GMFlow) with a motion-guided attention module, a feature separator, a two-branch module and a reconstruction network. For video motion magnification, given dense feature maps $F_a, F_b \in \mathbb{R}^{H\times W\times D}$ extracted from two consecutive video frames I_a, I_b, where H,W are the resolution size and D indicates the size of feature dimension, we aim to predict the magnified image Y and hence highlight the subtle motion signals. In the following, we describe our network architecture in detail.

Optical Flow. Based on dense feature maps F_a, F_b, we aim to extract optical flow by exploiting GMFlow [25]. Specifically, we compute the matching distribution $\mathcal{C}$ of the correlations between feature maps with a softmax function,

$$\mathcal{C} = \text{softmax}\left(\frac{F_a F_b^T}{\sqrt{D}}\right) \in \mathbb{R}^{H\times W\times H\times W}. \tag{1}$$

Based on the matching distribution $\mathcal{C}$ and the 2D coordinates of pixel grid g, we can get optical flows $O = \mathcal{C}g - g$.

Motion-Guided Attention. To highlight the motion areas and reduce the annoying video artifacts, 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 $h(\cdot)$ [12] with a Sigmoid activation function as below:



(a) Lagrangian Branch (b) Eulerian Branch

Figure 2: Feature analysis. (a) A global attention map and local optical flow in our Lagrangian branch by using quantifying attention method [1]. (b) Global motion flow and a local activation map in our Eulerian branch by using the CAM method [27].

$$A = (\alpha - 1)\,\text{Sigmoid}(h(I, O)) + \mathbb{1}, \tag{2}$$

where α is the given magnification factor and $\mathbb{1}$ denotes an all-ones matrix of ones to retain optical flow estimation of the background and amplify the motion of the object.

Feature Separator and Reconstruction Module. For the feature separator, given the input frame I, it aims to get texture representations V and motion representations M through a three-layer residual block and two independent residual block heads. For the reconstruction module, it concatenates the re-scaled texture features and the motion features, and then generates the predicted magnified frames via convolutional layers and residual blocks.

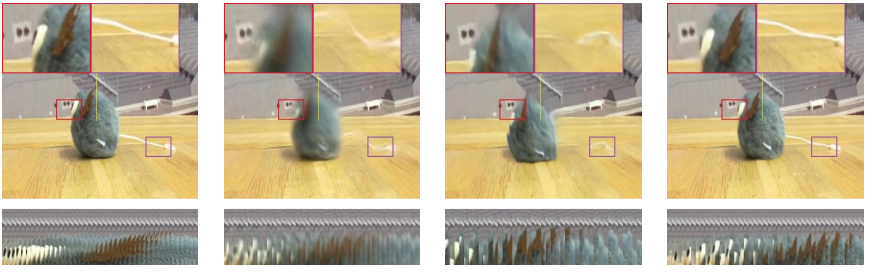
3.2 Two-Branch Module

We propose a sequence of L repeated two-branch module to combine global motion magnification quality from the Eulerian perspective with the precise magnification of moving objects by Lagrangian methods as shown in Fig. 1 bottom. Furthermore, our two-branch module with CNN and Transformer can maintain spatial correlations among adjacent frames, align the moving or jittering background and warp the magnified optical flow accurately. As shown in Fig. 2, for the Lagrangian branch, we capture global representations by Transformer to enlarge the attention range of optical flow magnification, which improves the ability to distinguish moving objects and the disturbing background. For the Eulerian branch, taking advantage of CNN's local reception fields, we first calculate the micro-movements of each pixel in the whole image and then target and magnify motion signals. In the following, we introduce two branches and their fusion process, respectively.

Lagrangian Branch. Our Lagrangian branch includes Transformer blocks that contain a global attention module (GA) [8], a manipulator that implements optical flow magnification and feedforward functions. Given the motion features of the first frame M_a, we generate query, key and value by three convolution operations and then obtain three independent patch embeddings. After the self-attention module and the folding operation to combine these patches, we apply another convolution layer to achieve the next dimensionally unified motion feature map M_{i+1}^{Lagra}, where i denotes the index of L two-branch modules as:

$$M_{i+1}^{Lagra} = \text{Res}[\Delta(\text{LN}(\bar{M}_i), A \odot O)]. \tag{3}$$

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



(a) Original video (b) VAM [26] (c) Oh *et al*. [16] (d) Ours

Figure 3: 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. From left to right, we show (a) original video, and the results from (b) VAM [26], (c) Oh *et al*. [16] and (d) our MagFormer.

module. We use bilinear interpolation $\Delta(\cdot,\cdot)$ for the optical flow warping. For the warping process, We have $\bar{M}_i = \mathrm{GA}(\mathrm{LN}(M_i)) + \mathrm{LN}(M_i)$.

Eulerian Branch. As shown in Fig. 1, each layer of the Eulerian branch is consist of several convolution blocks and a manipulator block that implements masked Eulerian magnification. For i^{th} layer, we have:

$$M_{i+1}^{Euler} = \mathrm{Conv}[M_b + \mathrm{Res}(\mathrm{Conv}(G(M_i) \odot A))], \tag{4}$$

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

Fusion. For the i^{th} layer of two-branch module, we get M_{i+1}^{Lagra} and M_{i+1}^{Euler} from Lagrangian branch and Eulerian branch based on Eqs. 3 and 4. We concatenate the features and then feed them into a fusion block, which is a stack of convolutional layers with the LeakyReLU activation, and finally get M_{i+1}.

3.3 Training

Inspired by Oh *et al*. [16], to construct the texture and the motion representations for input frames, we perturb the intensity of input frames by keeping the texture representations of perturbed frames to be the same, while their motion representation unchanged under perturbation. We adopt the robust Charbonnier loss function $\mathcal{L}_c$ between ground-truth magnified images Y and our predicted magnified images $\hat{Y}$ and also introduce losses based on our paired data constructions. Specifically, we encourage minimizing the distances between texture representations of consecutive frames V_a and V_b, between texture representations of input frames and perturbed frames V_b and V_b' and between motion representations of input frames and perturbed frames M_b and M_b'. We optimize the final objective:

$$\mathcal{L} = \mathcal{L}_c(Y,\hat{Y}) + \lambda(\mathcal{L}_c(V_a,V_b) + \mathcal{L}_c(V_b,V_b') + \mathcal{L}_c(M_b,M_b')), \tag{5}$$

where λ is a hyperparameter with a value of 0.1 in our paper.

PSNR/SSIM	Phase-based [21]	Oh *et al*. (Static) [16]	Oh *et al*. (Dynamic) [16]	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**

Table 1: 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 [21]	Oh *et al*. (Static) [16]	Oh *et al*. (Dynamic) [16]	Ours
Cat toy	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**

Table 2: Average PSNR and SSIM of different motion magnification methods of six videos with $\alpha = 40$. The presentation format is PSNR / SSIM. The best results are in bold.

4 Experiments

4.1 Implementation Details

Datasets. Our method is trained on a synthetic motion magnification dataset proposed by [16]. All the training images are 384×384 and the magnification factors range from 0 to 100. Considering the different advantages of Eulerian and Lagrangian methods, we select 3 background-fixed videos (eye, plants, and drum) and 3 constantly moving object videos (cat toy, drone, and bottle) from 9 videos in total (others are gun, guitar and baby) to balance the testing video set.

To further verify the effectiveness of our proposed method, we use a Nikon D7200 RGB camera to record the model vibration video of the modal exciter. Besides RGB sequences, the amplitude and frequency are carefully recorded for evaluation.

Evaluation Metrics. To evaluate the performance of different magnification methods, we adopt Peak Signal-to-noise Ratio (PSNR) and Structural Similarity Index (SSIM) to evaluate the quality of magnified images. Also, we propose to measure the performance of motion magnification based on the motion amplitude and frequency as magnifying the motions should not change the natural frequency.

Experimental Settings. We use ADAM [10] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ for optimization. The total iteration is set to be 360,000 with a batch size of 4. We use cosine annealing [14] to update the learning rate, where the initial learning rate is 2×10^{-4} and the minimum learning rate is 10^{-7}. All the models are trained on 4 NVIDIA TITAN XP GPUs.

4.2 Results on Real-World Videos

In this section, we compare our method with state-of-the-art motion magnification methods [16, 21, 26]. We choose the magnification factors as 10, 20, and 40 since the motion magnification methods are intended to amplify subtle motions at high magnifications.

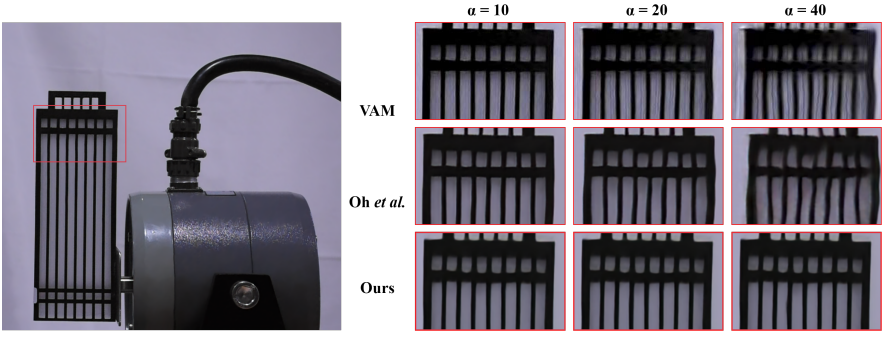


Figure 4: Comparison with VAM [26] and Oh *et al*. [16] on the exciter videos with different α.

Quantitative Results. We show the average PSNR and SSIM of all testing videos with different α in Table 1. As presented, our method outperforms the phase-based method [21] and the learning-based static method [16] significantly. Compared to the learning-based dynamic method [16], our method achieves comparable performance when the magnification factor is relatively low ($\alpha = 10$). With the magnification factor up to 20, our method achieves the best results, including an impressive 1.33 dB and 0.0152 increase in PSNR and SSIM, respectively. Moreover, the improvement becomes more obvious when $\alpha = 40$.

To verify the image quality in the high magnification factor, we calculate the PSNR and SSIM of different videos with $\alpha = 40$ as shown in Table 2. For the constantly moving objects videos (including the cat toy, drone, bottle), our MagFormer achieves the best performance both in PSNR and SSIM as the motion magnification attention of constantly moving objects is easy to obtain. For the background-fixed subtle motion videos (including eye, plants, drum), our MagFormer outperforms in PSNR and SSIM except the eye video.

Qualitative Results. We compare our proposed method with state-of-the-art methods [16, 26], and show the qualitative results. We can see the upper part of the stack single column of pixels of acceleration [26] and learning-based method [16] is horizontally uneven and their two sub-regions are also excessive blurring, as the Eulerian method's inability to choose motions of interest. In contrast, ours achieves superior performance thanks to the benefits of the two-branch module and the motion magnification attention that our method can pick the trajectories of moving objects and magnify the motion of interest. As we can see, the lower part of the stack column of our MagFormer shows the magnified motion of the cat toy, while the upper part and the two selected regions (which represent the background with slight vibration) are highly similar to the original video.

4.3 Results on Vibration Videos

Here, we verify the motion amplitude and frequency of input videos after motion magnification. We fix a building model on a modal exciter and control the model to repeat the sinusoidal vibration at a constant amplitude and frequency. We record the amplitude and frequency from a laser displacement sensor as the ground truth and evaluate our original videos and magnified videos.

Qualitative Results. As shown in Fig. 4, our MagFormer can generate more refined details when retaining physically accurate frequencies and amplitudes. In contrast, with the rising

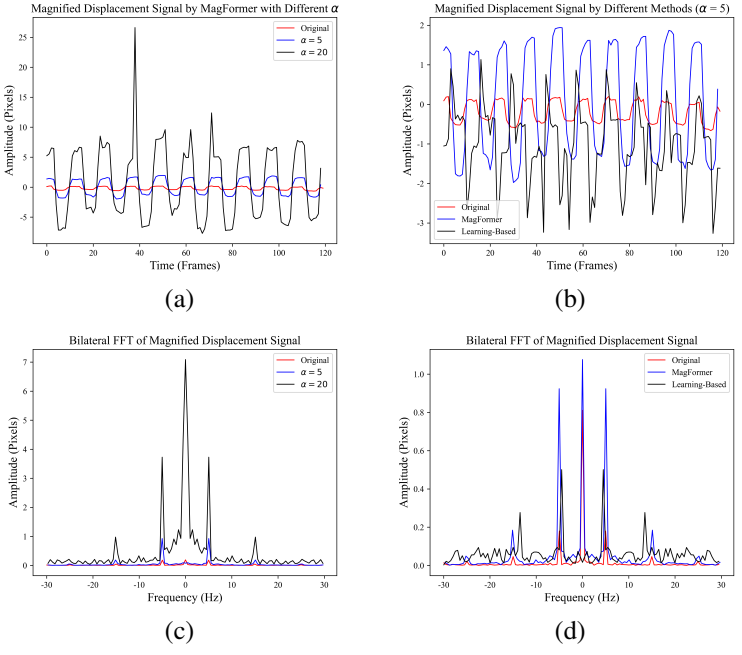


Figure 5: Amplitude and frequency of motion signal detected using different magnification methods and magnification factors. Magnified displacement signal of (a) MagFormer with different α and (b) different methods. Bilateral FFT of magnified displacement signal of (c) MagFormer with different α and (d) different methods.

magnification factor α, VAM [26] and learning-based Eulerian method [16] fail to handle sharp horizontal and vertical building patterns. Meanwhile, the ringing effect of VAM disturbs the detection of frequencies and amplitudes. These results also verify the superiority of our MagFormer on texture generations.

Quantitative Results. We measure the amplitude of the magnified signal of MagFormer with different magnification factors. As shown in Fig. 5(a), when the maximum amplitude of the original video is 0.86 pixels, the maximum amplitude of magnification video with 5 times and 20 times magnification is 3.91 pixels and 17.32 pixels respectively. Also, when the motion frequency is 5 Hz, the detected frequency of the original video, 5 times magnification video, and 20 times magnification video is 4.9950 Hz, 5.0370 Hz, and 5.0370 Hz respectively, and the relative errors are all less than 0.8% (See Fig. 5(c)). This verifies that our MagFormer can keep the proposition of motions unchanged during the magnification procedure. For the frequency consistency and amplitude of the output signal, we compare MagFormer with the learning-based method [16] with $\alpha = 5$. As shown in Fig. 5(d), while the motion frequency of the original video is 4.9950 Hz, the detected frequency of Magformer's result and [16]'s result are 5.0370 Hz and 4.4955 Hz respectively. Our MagFormer achieves a lower relative error in the frequency (0.8% versus 10%). Also, the spectrum results are MagFormer are similar to the original signal, while the spectrum of the learning-based method shows inward contraction. As for Fig. 5(b), we can see that our MagFormer can keep the magnified signal's waveform similar to the original, while the waveform of

Two-branch module	Layer-by-layer	Attention	PSNR
Two-branch	w/	w/	30.28
Two-branch	w/	w/o	29.61
Two-branch	w/o	w/o	28.40
Lagrangian only	w/	w/o	25.56
Eulerian only	w	w/o	24.57

Table 3: Comparison of MagFormer with different branches, with (w/) and without (w/o) the layer-by-layer manner in Lagrangian branch and motion magnification attention.

the learning-based method's result changes. In conclusion, our MagFormer can magnify the motion signal and keep the spectrum of motions unchanged.

4.4 Ablation Study

We conduct an ablation study on selected seven videos to analyze the performance of attention, and different network structures in Table 3, respectively. For the motion magnification attention, we compare training with and without motion magnification attention while setting the network structure as proposed two-branch. We can see that training with attention can achieve an increase in PSNR of 0.67 dB. To show the superiority of our two-branch module, we also compare our two-branch module with only Lagrangian branch (*i.e.*, Lagrangian only) and only Eulerian branch (*i.e.*, Eulerian only). The usage of the two-branch module outperforms both Eulerian only and Lagrangian only, with increases in PSNR of 5.04 dB and 4.05 dB. To verify the importance of the layer-by-layer manner, we also design two-branch where the iteration only occurs in Eulerian branch. Despite a slight decrease compared to the original two-branch module (-1.21 dB), it still increases PSNR compared to the one-branch structure (3.83 dB and 2.84 dB). Hence, we find out that our two-branch design is the main contributor to the improvement in image quality.

5 Conclusion

In this work, we propose a novel unified framework, MagFormer, for video motion magnification. The framework integrates global motion feature flow and local moving object optical flow and magnifies them through a layer-by-layer pattern. Especially, we introduce a hybrid two-branch module with a Transformer branch from Eulerian perspective and a CNN branch from Lagrangian perspective. Compared to prior state-of-the-art methods, we show that our method has less ringing effect and retains high-quality texture feature up to a higher magnification factor. Moreover, we introduce a new vibration dataset and a corresponding metric to evaluate video motion magnification quantitatively via amplitude and frequency. The experimental results demonstrate the effectiveness of our magnification method in terms of texture generation quality as well as precisely preserving the original physical properties.

6 Acknowledgements

This work was supported by "the Fundamental Research Funds for the Central Universities", and the National Natural Science Foundation of China under Grant 62076016, Beijing Natural Science Foundation-Xiaomi Innovation Joint Fund L223024.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, 2020.

[2] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Mohamed Elgharib, Mohamed Hefeeda, Fredo Durand, and William T Freeman. Video magnification in presence of large motions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4119–4127, 2015.

[5] Vincenzo Fioriti, Ivan Roselli, Angelo Tatì, Roberto Romano, and Gerardo De Canio. Motion magnification analysis for structural monitoring of ancient constructions. *Measurement*, 129:375–380, 2018.

[6] Philipp Flotho, Mayur J Bhamborae, Lars Haab, and Daniel J Strauss. Lagrangian motion magnification revisited: Continuous, magnitude driven motion scaling for psychophysiological experiments. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3586–3589. IEEE, 2018.

[7] Philipp Flotho, Cosmas Heiss, Gabriele Steidl, and Daniel J Strauss. Lagrangian motion magnification with double sparse optical flow decomposition. *arXiv preprint arXiv:2204.07636*, 2022.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[9] Mirek Janatka, Hani J Marcus, Neil L Dorward, and Danail Stoyanov. Surgical video motion magnification with suppression of instrument artefacts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 353–363. Springer, 2020.

[10] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[12] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7274–7283, 2019.

[13] C. Liu, A. Torralba, W. Freeman, F Durand, and E. Adelson. Motion magnification. *ACM Transactions on Graphics*, 24(3):p. 519–526, 2005.

[14] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[15] Mingyuan Mao, Renrui Zhang, Honghui Zheng, Teli Ma, Yan Peng, Errui Ding, Baochang Zhang, Shumin Han, et al. Dual-stream network for visual recognition. *Advances in Neural Information Processing Systems*, 34:25346–25358, 2021.

[16] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Fr'edo Durand, William T Freeman, and Wojciech Matusik. Learning-based video motion magnification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 633–648, 2018.

[17] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 367–376, 2021.

[18] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[19] Shoichiro Takeda, Kenta Niwa, Mariko Isogawa, Shinya Shimizu, Kazuki Okami, and Yushi Aono. Bilateral video magnification filter. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17369–17378, 2022.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[21] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013.

[22] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Riesz pyramids for fast phase-based video magnification. In *2014 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2014.

[23] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (TOG)*, 31(4):1–8, 2012.

[24] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.

[25] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022.

[26] Yichao Zhang, Silvia L Pintea, and Jan C Van Gemert. Video acceleration magnification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 529–537, 2017.

[27] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.