Supplementary Materials to "MEGAN: Memory Enhanced Graph Attention Network for Space-Time Video Super-Resolution"

Chenyu You[1] Lianyi Han[2] Aosong Feng[1] Ruihan Zhao[3]
Hui Tang[2] Wei Fan[2]
[1]Yale University [2]Tencent Hippocrates Research Lab
[3]The University of Texas at Austin

In this supplementary material, we provide further comparisons to investigate the effects of different components in MEGAN.

1. More Ablation Study

1.1. Further Analysis on Different Components

We conduct experiments of MEGAN w/o GCN, MEGAN w/o non-local, MEGAN w/o global-local feature aggregation (GL-Agg). In the table 1, results show that the proposed method *achieved significant improvements* over i). w/o GCN, ii). w/o NLRB, or iii). w/o GL-Agg. This proved the effectiveness of each component.

Method	Metric	w/o GCN	w/o NLRB	w/o GL-Agg	MEGAN (our)
Vid4	PSNR	26.42	26.44	26.45	**26.57**
	SSIM	0.8007	0.8017	0.8020	**0.8044**
Vimeo-Fast	PSNR	36.93	37.01	37.03	**37.18**
	SSIM	0.9424	0.9430	0.9431	**0.9446**

Table 1: Ablation study on different components over Vid4 and Vimeo-Fast dataset.

1.2. Further Analysis on NLRB

The major differences in NLRB between [4] and our method includes: i) [4] uses Gaussian function as the pairwise function, but we use dot product similarity. The reason why we use dot product similarity is to boost convergence, which allows us to train very deep networks, being more suitable for STVSR. Table 2 reports the performance of using three types of functions in 600,000 iterations on 2 1080Ti GPUs; 2) Using [4] introduces higher computational complexity, especially in the context of limited GPU resources.

1.3. NLRB vs LMGA

In our work, NLRB is used to capture channel correlations among low-level features because it utilizes information in a short range. Since the enlarged (interpolated) features are often of low quality, LMGA is designed to refine interpolated frame features for better spatial alignment, and model temporal correlations among whole videos because the information from longer content could be utilized. By well exploiting both short and long-range space-time dependencies, it will complementarily enhance performance. The results in Table 1 also validate our claims.

Method	Gaussian function	Embedded Gaussian function	Dot Product Similarity (our)
PSNR	26.46	26.39	**26.57**
SSIM	0.8024	0.8001	**0.8044**

Table 2: Ablation study on different components over Vid4 and Vimeo-Fast dataset.

1.4. Random Sampling in LMGA

The feature aggregation in LMGA module is designed to fuse local feature and global feature, which is sampled from the shuffled global pool to capture the global information. We interpret random sampling as an effective data augmentation: the key frame feature can gather information from different subsets of frame features among the same video in different training steps, which further improves MEGAN's generalization ability. In Table 3, we also compare random sampling with uniform sampling on Vid4, proving the effectiveness of random sampling. This is because uniform sample only captures a short local range of semantics, while random sampling can utilize rich information beyond a fixed content. We also can observe that "w/o GL-Agg" by removing feature aggregation results in the performance degradation in Table 1.

Method	PSNR	SSIM
Uniform Sampling	26.45	0.8018
Random Sampling (our)	**26.57**	**0.8044**

Table 3: Ablation study on random sampling in LMGA.

Table 4: Ablation study of different components on Vid4 [2].

Case Index	1	2	3	4	5	6	7
Channel-based Attention Block [5]	✓	✗	✗	✗	✓	✗	✗
Non-local Block [3]	✗	✓	✗	✗	✗	✓	✗
Non-local Residual Module [4]	✗	✗	✓	✗	✗	✗	✓
LMGA	✗	✗	✗	✓	✓	✓	✓
PSNR (dB)	26.27	26.34	26.40	26.44	26.47	26.51	26.57
SSIM	0.7985	0.7991	0.7999	0.0817	0.8018	0.8021	0.8044

Table 5: Ablation study of block number on Vid4 [2].

Case Index	1	2	3	4	5	6	7	8	9	10
PFRDB Number	0	5	10	20	30	5	5	5	5	5
ResBlock Number	20	20	20	20	20	0	5	10	15	30
PSNR (dB)	26.45	26.57	26.57	26.58	26.58	26.39	26.46	26.50	26.54	26.57
SSIM	0.8024	0.8044	0.8045	0.8047	0.8045	0.8008	0.8012	0.8021	0.8028	0.8045
Parameter Number	10.01M	10.71M	11.41M	12.80M	14.20M	9.23M	9.60M	9.97M	10.34M	11.46M

1.5. Further Analysis on Attention-based Blocks and LMGA

In Table 4, when comparing cases 1-3, we use different types of attention blocks (Channel-based Attention Block [5], Non-local Block [3], and Non-local Resblock [4]) instead of the proposed LMGA module. We can observe that the non-local module brings performance improvements. This suggests that, by exploiting low-level and high-level features, MEGAN is capable of utilizing non-local information to learn better representational ability. In case 4, we also learn that the proposed LMGA block contributes to superior performance gains, regardless of whether we use attention-based blocks (cases 1-3). This demonstrates that we can significantly improve the reconstruction performance in space-time domain by dynamically incorporating information from spatial features and temporal contexts.

We fix the proposed LMGA in MEGAN in cases 5-7. As seen in Table 4, we observe that performance improvements benefit from utilizing Non-local Resblock, suggesting adopting residual non-local learning is able to capture long-range spatio-temporal correlations, which results in a marginal gain in network performance. Besides, residual learning [1] is employed to promote the training process more stable.

1.6. Further Analysis on Block Number

In this section, we conduct extensive experiments to investigate the effects of block number in Table 5. When comparing cases 1-5, we investigate the effects of PFRDB while fixing the number of ResBlock at 20. We observe that adding more PFRDBs [4] achieve a small but consistent performance improvement. However, introduction of PFRDBs requires more computational cost and much training time. So we use 5 PFRDBs to learn more spatial-temporal information by leveraging low- and high-level features.

Besides, we compare cases 6-10 to study the effects of ResBlock when the number of ResBlock is fixed at 5. It can show that we achieve a marginal performance gain when adding more ResBlocks. This suggests that we can achieve robust representation ability while maintaining relatively small network size. To balance the trade-off between computational cost and time, we adopt 20 ResBlocks here.

Overall, our MEGAN can make superior improvements including 5 PFRDBs and 20 ResBlocks. Using the proposed LMGA module, our MEGAN can better handle complex dynamic space-time scenes

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[2] Ce Liu and Deqing Sun. A bayesian approach to adaptive video super resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 209–216. IEEE, 2011.

[3] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7794–7803, 2018.

[4] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Int. Conf. Comput. Vis.*, pages 3106–3115, 2019.

[5] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018.