1.5em 0pt

EE837 Term Project Final Report
Meta-Transfer Learning for Light Field Super-resolution

Sojeong Song

Abstract

In the recent studies, convolutional neural networks (CNNs) have been used dominantly in light field super-resolution, and they have brought dramatic advances in performance by using large-scale external datasets. Although exhaustive training on the external dataset has outperformed traditional light field super-resolution methods, deep learning-based methods have suffered from domain shift problems leading to greater performance degradation, especially in light field super-resolution where both spatial and angular correspondence has to be learned. To address this problem, light field zero-shot super-resolution (LFZSSR) has been introduced for self-supervised learning to avoid domain shift issues. However, training data is highly limited under zero-shot setting, so it is difficult to train an end-to-end network. Moreover, LFZSSR requires thousands of gradient updates because of the limited training data, so real-time application is not available because of long inference time. To overcome these limitations of LFZSSR, in this paper, we tried to search a novel architecture for Meta-Transfer Learning in Light Field Super-Resolution (LFMZSR), which inspired by meta-transfer learning for zero-shot super-resolution (MZSR). Meta-transfer learning finds a base-learner, which makes few gradient updates yield competitive performances compared to the existing state-of-the-art methods. Experiments on many variations of neural network structures for view alignment and multi-view aggregation were conducted. The limited training data could overcome with neural architecture search for light field super-resolution and pretrained model with external large-scale datasets. Our model was evaluated on a variety of real-world and synthetic datasets, and it showed potential for significant improvement in terms of inference time as well as performance with no domain gap problem. Future work will use our model to boost performance for real-time practical applications.

1. Introduction

A 4D light field, which has spatial and angular information of light in space, has gotten more attraction to the researchers since a portable light field camera is available. A light field can be obtained by an array of cameras and displayed by a display panel with a lens array such as a pinhole array. Although micro-lens-array cameras have developed and commercialized, there is a trade-off between spatial and angular resolution. As the number of lenses in the array have increased to get more angular information, the size of the lenses has gotten smaller resulting in the reduction of spatial resolution of light field. The degradation of spatial resolution restricts the performance of future applications such as depth estimation and rendering.

Super-resolution (SR) has been considered one of the most important topics to enhance the spatial resolution of reference views and actively studied (Shi et al., 2020). In the recent studies, CNN-based techniques are dominantly used to super-resolve reference views of a light field exploiting the power of deep learning and demonstrate significant PSNR(dB) and SSIM compared to the classical light field SR methods (Yoon et al., 2015). The recent SotA CNN-based SR models require a long time and have overhead on memory since they are based on exhaustive training with external light field dataset. Although they outperform the classic methods with a large margin, they are not input-specific so the performance may not ensure when the domain gap is large. Domain gap refers to the gap between real-world data and synthetic data, and when the trained model is applied to real-world data, it causes the performance drop known as domain shift problem.

Light field SR with zero-shot learning (LFZSSR) addressed the domain shift problem by implicitly learning across-scale internal recurrence in an unsupervised or self-supervised manner (Cheng et al., 2021). LFZSSR inspired by zero-shot single image SR (ZSSR) and trained a small and simple CNN networks for view alignment and multi-view aggregation exploiting a low-resolution (LR) light field and its downscaled version (Shocher et al., 2017). LFZSSR achieved noticeable improvements in light field SR performance, especially when the domain gap was large.

Even though LFZSSR showed incredible results without the domain shift problem, there was a few limitations. First, the zero-shot framework took time to get high-resolution (HR) results since it required thousands of gradient updates at inference time. Moreover, it could not take the advantages of end-to-end networks due to the limited

training data, so the learning process was complex and hand-designed. These limitations make it difficult to apply the model on embedding systems and automatically use the model as real-time applications.

Just as LFZSSR was inspired by ZSSR, we propose a meta-transfer learning framework for light field SR inspired by meta-transfer learning for zero-shot super-resolution (MZSR) (Soh et al., 2020). Meta-transfer learning aims to initialize the gradient exploiting a large-scale synthetic dataset and learn a base-learner to move the initialized gradient point where it can quickly adapt to a new task with a few gradient updates. We found that only few gradient updates are sufficient to find a suitable gradient value, and the proposed framework outperforms LFZSSR in much shorter inference times. The proposed framework extracted common representations from large-scale datasets such as SAE dataset, exploiting both internal and external information. Different reference views were assumed to be different kernels in the meta learning algorithm to train the base-learner, and it quickly found a high-quality HR view regardless of the reference view. However, meta learning was sensitive to the structure of neural networks so the results were not stable. Therefore, we conducted additional experiments for proper architecture search at view alignment stage and multi-aggregation stage. The limited training data could be overcome with architecture search and meta learning with reference views.

2. Related work

2.1. Classical light field SR

The light field can be obtained from lens arrays, and a raw light field is a lenslet image. The raw image is easily transformed to a sub-aperture image by simply moving of pixels. Light field SR aims to find HR images of reference views and classical light field methods estimates depth in images by calculating disparities trying to figure out the structure of 3D objects. Rossi and Frossard proposed mathematical modeling based on graph theory to find the structure of light field (Rossi & Frossard, 2018). Wanner and Goldluecke optimized the reconstruction of disparity by using geometric property such as epipolar plane to generate super-resolved views (Wanner & Goldluecke, 2014). CNN-based light field SR shows steadily higher performance than the SotA classical light field SR methods.

2.2. CNN-based light field SR

Numerous CNN-based methods have demonstrated their ability to internally learn representations of light fields extracted from external datasets. Yoon proposed LFCNN which first adopt a deep convolutional neural network (CNN) to solve image restoration problems (Yoon et al., 2015). Meng formulated light field SR as tensor restoration and designed a stage-wise loss function based on feature maps (Meng et al., 2021). Most studies proposed diverse network structures, and they show notable performance improvements but they severely suffer from the domain shift problem. In contrast, LFZSSR is self-supervised learning so it is domain-agnostic.

2.3. LFZSSR

Self-supervised learning method uses only one test image to train the network so it can avoid the domain shift problem by preventing the overfitting problem on training dataset. Cheng et al. proposed the framework for light field SR using self-supervised learning method that automatically learned internal recurrence of light field (Cheng et al., 2021). Even though the model showed better adaptation in real-world scenarios than state-of-the-art deep learning-based methods and boosted the performance of SR, there were two limitations. In this paper, we revise the model and propose a novel framework that complements the weaknesses and makes use of the strengths of the LFZSSR.

2.4. Meta learning

Meta learning aims to design the model that can quickly adapt to a new task with a few iterations. Generally, there are three approaches in meta learning. Metric based methods learn efficient distance metric while memory network based methods usually use recurrent networks exploiting external and internal memory. Optimization based methods aims to optimize parameters of the model by fast learning. One of commonly used for optimization based meta learning methods is Model-Agnostic Meta-Learning (MAML), and the algorithm was revised to learn a base-learner for SR of reference views (Finn et al., 2017).

3. Method

The overall flow of the proposed algorithm can be seen in the Algorithm 1, and the main framework is based on the LFZSSR's framework. Preupsampling, view alignment, and multi-view aggregation are three stages for light field SR self-supervised learning. Preupsampling processed with a simple VDSR, which has 18 residual layers, is used for estimating disparity when wrapping. View alignment conducts disparity-guided warping, and multi-view aggregation super-resolves the central view with the wrapped light field that gives the neighboring scenes information to the central view.

3.1. View alignment

The reference views of the sub-aperture image contain rich information about subjects viewed from different

Algorithm 1 Light Field Meta-Transfer Learning

Data: LLR light field Z^{LLR} and LR light field Z^{LR}
Data: learning rate α, β
Result: Model parameter θ_M
Randomly initialize θ_{align} and θ_{aggre}
while *not done* **do**
  Sample LR patch from large-scale dataset
  Update θ_{align} with respect to L_{align} by Eq.1
end
Fix the AlignNet parameters
while *not done* **do**
  Sample LLR-LR pair
  while *not done* **do**
    Input reference views as tasks T_i
    Calculate gradient descent of training loss by Eq.4
  end
  Update θ_{aggre} with respect to L_{aggre} by Eq.5
end

angles. As long as the information spread across the different reference views is gathered properly, it is a great source to improve the quality of super-resolution of central view. The geometry correspondences can be matched by view alignment. View alignment has three steps processing with patches extracted from target LR light field. First, a plane-sweep volume (PSV) generated from a patch allows us to find the disparity map more efficiently, and PSV has been widely used for classical disparity estimation methods. Then, PSV is fed into alignment-oriented disparity estimation network(AlignNet). AlignNet consists of several convolutional layers, for example, AlignNet could be constructed with 5 convolutional layers, one for feature extraction and the other for disparity estimation as like LFZSSR. Disparity map generated from AlignNet is up-scaled to the target resolution and wrapped with the preupsampled central view. Then, each view of wrapped light field should be the same with the preupsampled central view. Therefore, the alignment loss function is defined as

$$L_{align} = \frac{1}{N}\sum_{n=1}^{N}\sum_{u\in U}\left\|W_n^{LR}[u] - P_n^{LR}\left[u_c\right]\right\|_2^2. \quad (1)$$

The parameter of view alignment can be initialized by the parameter of pretrained model made with a large dataset such as HFUT or SAE.

3.2. Multi-view aggregation

The wrapped light field is used to train a multi-view aggregation network (AggreNet). To avoid the domain gap problem, we generate LLR-LR training pairs by Bicubic downsampling LR light field patches. The LLR and LR light field patches are each aligned via AlignNet with fixed parameters, and the resulting wrapped light field pairs are used to train AggreNet. Here, 5-layers CNN is used to super-resolve the central view image by exploiting neighboring scenes information. The super-resolved LLR central view should be the same with the LR central view. Therefore, the recurrence loss function is defined as

$$L_{recur} = \frac{1}{N}\sum_{n=1}^{N}\left\|S_n^{uR}\left[u_c\right] - Z_n^{LR}\left[u_c\right]\right\|_1. \quad (2)$$

In order to avoid missing high-frequency information, additional back-projection loss is defined as

$$L_{bp} = \frac{1}{N}\sum_{n=1}^{N}\left\|S_n^{LR}\left[u_c\right]\downarrow_\alpha - Z_n^{LR}\left[u_c\right]\right\|_1. \quad (3)$$

Pretrained model for multi-aggregation is generated by using a large light field dataset HFUT or SAE, and from the gradient point, the gradient is optimized by meta learning. The reference views are used as support sets for meta learning, and each gradient descent for training loss is calculated by the equation as

$$\theta_i = \theta - \alpha\nabla_\theta L_{T_i}^{t_r}(\theta). \quad (4)$$

The full gradient is updated by average test loss which is defined as

$$\theta \leftarrow \theta - \beta\nabla_\theta\Sigma_{T_i}L_{T_i}^{te}\left(\theta_i\right). \quad (5)$$

To sum up, the total loss function is described as

$$L_{aggre} = L_{recur} + \gamma_1 L_{bp}. \quad (6)$$

$$L_{total} = L_{aggre} + \gamma_2 L_{align}. \quad (7)$$

4. Experimental results

4.1. Dataset

A real-world HFUT dataset with 640 scenes and a synthetic SAE dataset with 180 scenes are used as large-scale source datasets for transfer learning or testing. For target datasets, real-world datasets such as EPFL dataset with 20 scenes and Stanford Lytro Archive (Stan) dataset with 20 scenes and synthetic datasets such as HCI1 dataset (old) with 10 scenes and HCI2 dataset (new) with 20 scenes are used to validate the flexibility of the proposed model facing the domain shift problem. For real-world datasets, light fields are cropped by 9x9 central views to avoid the vignetting effect. Experiments were conducted in a zero-shot setting where only one image is available at test time. For evaluation, the most widely used measuring tool, peak signal to noise ratio (PSNR), was used to estimate the quality of HR light fields.

4.2. Evaluation

As shown in Table 1 and 2, the PSNR results of our proposed method, LFMZSR, are compared with the previous super-resolution method when the data distribution of a source dataset is far from the distribution of a target dataset. HFUT dataset and SAE dataset are used to pretrain the AlignNet and AggreNet. The scale of super-resolution was 2, and the learning rates for wrapping, fusion, finetuning were 1e-5. These are the results of LFZSSR, LFMZSR with 10 iterations, and LFMZSR with 100 iterations at the test time. The improvement can be seen over all datasets, which proves that the proposed LFMZSR is the most effective to learn the internal recurrence even when spatial and angular correspondences construct highly complex hidden structures. These results validate that our method preserves low-frequency information and restores the high-frequency details. The redundancies in the reference views of the light field are aggregated to the central view of the light field and gives high-frequency information. LFZSSR achieved the results by updating the gradients 20000 times and one update took about 450 seconds, so it took about 2 hours to get one super-resolved central image. On the other hand, our method has comparable performances with much less iterations at the inference time and shows the best performance beyond the previous LFZSSR method with one hundred iterations. This implies that real-time applications are possible by boosting the performance using our framework.

Table 3 shows the PSNR results of super-resolution when changing CNN architectures of AlignNet and AggreNet. Other hyperparameter setting was the same with upper experiment. LFZSSR used a simple structure for AlignNet and AggreNet, and there was no experiment for searching appropriate architecture. To find optimal structure for each network, we evaluated the performance with various CNN architectures with different depths. As can be seen, 10-layers CNN showed the higher performance than the previous setting, and 100-layers CNN showed good performance on some datasets. This shows the optimization of the alignment network and the multi-view aggregation network is needed, and it can result the significant improvement on the performance of light field super-resolution.

Table 1. The average PSNR results of light field super-resolution on various target datasets when using the pretrained model trained on HFUT dataset.

	LFZSSR	LFMZSR-10	LFMZSR-100
EPFL	32.6732	32.2304	32.8806
STANFORD	31.8001	31.5543	32.4966
HCI1	43.6142	43.5385	43.8229
HCI2	32.1466	32.0082	32.4986

Table 2. The average PSNR results of light field super-resolution on various target datasets when using the pretrained model trained on SAE dataset.

	LFZSSR	LFMZSR-10	LFMZSR-100
EPFL	33.2764	33.2849	33.7831
STANFORD	32.8321	32.3884	33.9267
HCI1	43.2324	41.2378	42.9281
HCI2	34.9822	33.4345	34.9203

5. Discussion

In this project, we showed LFMZSR can reduce the inference time significantly keeping the performance of super-resolution. However, we found the two points to discuss about. First, many hyperparameters have to be tuned because the framework is still complex. So the future studies for simplification of the framework may reduce the redundant time for tuning. Moreover, we found that there is a room to optimize the CNN architectures to improve the performance of super-resolution. Second, meta learning is sensitive to the choice of learning scheduling including setting hyperparameters. The performances were not stable, and the meta learning highly depends on the neural network structures. Therefore, the future work should use additional algorithms to stabilize the learning and search for the best structures.

References

Cheng, Z., Xiong, Z., Chen, C., Liu, D., and Zha, Z.-J. Light field super-resolution with zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10010–10019, June 2021.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks, 2017.

Meng, N., So, H. K.-H., Sun, X., and Lam, E. Y. High-dimensional dense residual convolutional neural network for light field reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):873–886, Mar 2021. ISSN 1939-3539. doi: 10.1109/tpami.2019.2945027. URL http://dx.doi.org/10.1109/TPAMI.2019.2945027.

Rossi, M. and Frossard, P. Geometry-consistent light field super-resolution via graph-based regularization. *IEEE Transactions on Image Processing*, 27(9):4207–4218, Sep 2018. ISSN 1941-0042. doi: 10.1109/tip.2018.2828983. URL http://dx.doi.org/10.1109/TIP.2018.2828983.

Shi, J., Jiang, X., and Guillemot, C. Learning fused pixel and feature-based view reconstructions for light fields. In

Table 3. The average PSNR results of light field super-resolution on various target datasets when changing the depth of the CNN architectures for AlignNet and AggreNet.

	5-LAYERS	10-LAYERS	50-LAYERS
EPFL	32.6732	32.7212	32.3244
STANFORD	31.8001	31.6705	30.2232
HCI1	43.6142	43.7230	42.6463
HCI2	32.1466	32.2213	31.9532

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

Shocher, A., Cohen, N., and Irani, M. "zero-shot" super-resolution using deep internal learning, 2017.

Soh, J. W., Cho, S., and Cho, N. I. Meta-transfer learning for zero-shot super-resolution, 2020.

Wanner, S. and Goldluecke, B. Variational light field analysis for disparity estimation and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):606–619, 2014. doi: 10.1109/TPAMI.2013.147.

Yoon, Y., Jeon, H.-G., Yoo, D., Lee, J.-Y., and Kweon, I. S. Learning a deep convolutional network for light-field image super-resolution. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 57–65, 2015. doi: 10.1109/ICCVW.2015.17.