Detail-Preserving Transformer for Light Field Image Super-Resolution

Shunzhou Wang¹, Tianfei Zhou²*, Yao Lu¹*, Huijun Di¹
¹ Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, China
² Computer Vision Laboratory, ETH Zurich, Switzerland
{shunzhouwang, vis_xl, ajon}@bit.edu.cn tianfei.zhou@vision.ee.ethz.ch

Abstract

Recently, numerous algorithms have been developed to tackle the problem of light field super-resolution (LFSR), i.e., super-resolving low-resolution light fields to gain high-resolution views. Despite delivering encouraging results, these approaches are all convolution-based, and are naturally weak in global relation modeling of sub-aperture images necessarily to characterize the inherent structure of light fields. In this paper, we put forth a novel formulation built upon Transformers, by treating LFSR as a sequence-to-sequence reconstruction task. In particular, our model regards sub-aperture images of each vertical or horizontal angular view as a sequence, and establishes long-range geometric dependencies within each sequence via a spatial-angular locally-enhanced self-attention layer, which maintains the locality of each sub-aperture image as well. Additionally, to better recover image details, we propose a detail-preserving Transformer (termed as DPT), by leveraging gradient maps of light field to guide the sequence learning. DPT consists of two branches, with each associated with a Transformer for learning from an original or gradient image sequence. The two branches are finally fused to obtain comprehensive feature representations for reconstruction. Evaluations are conducted on a number of light field datasets, including real-world scenes and synthetic data. The proposed method achieves superior performance comparing with other state-of-the-art schemes. Our code is publicly available at: https://github.com/BITszwang/DPT.

Introduction

Light field (LF) imaging systems offer powerful capabilities to capture the 3D information of a scene, and thus enable a variety of applications going from photo-realistic image-based rendering to vision applications such as depth sensing, refocusing, or saliency detection. However, current light field cameras naturally face a trade-off between the angular and spatial resolution, that is, a camera capturing views with a high angular sampling typically at the expense of a limited spatial resolution, and vice versa. This limits the practical applications of LF, and also motivates many efforts to study super-resolution along the angular dimension (i.e., to synthesize new views) or the spatial dimension (i.e., to increase the spatial resolution). Our work focuses on the latter one.

*Corresponding author

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
resolution tasks, mainly because 1) they represent each input image with small-size (e.g., $16 \times 16$ or $32 \times 32$) patches (i.e., tokens), which may damage fundamental structures (e.g., edges, corners or lines) in images, and 2) the vanilla fully connected self-attention design focus on establishing long-range dependencies among tokens, however, ignoring the locality in the spatial dimension. To address these problems, we introduce spatial-angular locally-enhanced self-attention (SA-LSA), which strengthens locality using convolutions first, then promotes non-local spatial-angular dependencies of each SAI sequence.

Based on SA-LSA, we design a novel Transformer-based LFSR model, i.e., DPT, which simultaneously captures local structures within each SAI and global structures of all SAIs in the light field. Concretely, DPT consists of a content Transformer and a gradient Transformer to learn spatial-angular dependencies within each light field and the corresponding gradient images, respectively. It is also equipped with a cross-attention fusion Transformer to aggregate feature representations of the two branches, from which a high-resolution light field is reconstructed.

In a nutshell, our contributions are three-fold: 1) We reformulate the problem of LFSR from a sequence-to-sequence learning perspective, which is differentiated to prior works in a sense that it fully explores non-local contextual information among all sub-aperture images, better characterizing geometric structures of light fields; 2) We design a spatial-angular locally-enhanced self-attention layer, which, in comparison with its vanilla fully-connected counterparts, offers our Transformer a strong ability to maintain crucial local context within light fields; 3) We finally introduce DPT as a novel Transformer-based architecture, which not only mines non-local contexts from multiple views, but also preserving image details for each single view. Our DPT demonstrates the promising performance on multiple benchmarks, while maintaining similar network parameters and computational cost as existing convolution-based networks.

Related Work

Light Field Super-Resolution. Early deep learning-based methods usually respectively learn the spatial and angular information with two independent subnetworks: one subnetwork captures the spatial information and another one learns the angular information. For example, (Yoon et al. 2017) adopted the SRCNN (Dong et al. 2014) to separately process each SAI for spatial super-resolution and interpolated novel views for angular super-resolution with the angular super-resolution network. (Yuan, Cao, and Su 2018) utilized a single image super-resolution network to enlarge the spatial resolution of each SAI, and applied the proposed EPI enhancement network to restore the geometric consistency of different SAIs. Recently, many researchers seek to simultaneously capture the spatial and angular information with a unified framework. (Wang et al. 2018) built a horizontal and a vertical recurrent network to respectively super-resolve 3D LF data. (Zhang, Lin, and Sheng 2019) stacked the SAIs from different angular directions as inputs and sent them to a multi-branch network to capture the spatial and angular information. (Wang et al. 2021c) used the deformable convolution (Dai et al. 2017) to align and aggregate the center-view and surrounding-view features to conduct LFSR. (Wang et al. 2020) developed a spatial-angular interaction network to learn the spatial-angular information from the macro-pixel image constructed with different SAIs. However, a major drawback of these approaches is that they fail to consider the long-range dependency among multiple SAIs in learning rich spatial-angular representations. To address this issue, we propose a Transformer based LFSR model, in which three proposed Transformers are leveraged to establish the non-local relationship of different SAIs for more effective representation learning.

Transformer for Image Super-Resolution. Due to the powerful representative ability of Transformer, some recent works attempted to apply it for image super-resolution tasks, which significantly boost the performance and benefit relevant applications in computer vision (Zhou et al. 2020; Li et al. 2021; Wang et al. 2021a;b; Zhou et al. 2021). These works can be roughly grouped into two categories: Transformer for single image super-resolution and Transformer for multiple image super-resolution. The former one mainly uses the Transformer to mine the intra-frame long-range dependency for high quality image reconstruction, i.e., IPT (Chen et al. 2021), SwinIR (Liang et al. 2021), and ESRT (Lu et al. 2021). The latter one adopts the Transformer to explore the inter-frame context information for accurate image reconstruction, i.e., TTSR (Yang et al. 2020) for reference-based image super-resolution and VSR for video super-resolution (Cao et al. 2021). Motivated by these approaches, we devise the first Transformer based architecture for LFSR.

Our Approach

Problem Formulation. We treat the task of LFSR as a high-dimensional reconstruction problem, in which each LF is represented as a 2D angular collection of sub-aperture images (SAIs). Formally, considering an input low-resolution LF as $L_{LR} \in \mathbb{R}^{U \times V \times H \times W}$ with angular resolution of $U \times V$ and spatial resolution of $H \times W$, LFSR aims at reconstructing a super-resolved LF $L_{HR} \in \mathbb{R}^{U' \times V' \times H' \times W'}$ with $\alpha$ the upsampling factor. Following (Yeung et al. 2018; Zhang, Lin, and Sheng 2019; Wang et al. 2020, 2021c), we consider the case that SAIs distribute in a square array, i.e., $U = V = A$, where $A$ indicates the angular resolution along the horizontal or vertical direction.

Network Overview. The overall architecture of DPT is shown in Fig. 1. Given $L_{LR}$ as the network input, we compute a gradient map for each 2D SAI, and organize them together as a gradient field $G_{LR} \in \mathbb{R}^{A \times A \times H \times W}$. $L_{LR}$ and $G_{LR}$ are separately fed into two small CNNs for convolutional feature extraction, following by two unimodal Transformers (i.e., a content Transformer and a gradient Transformer) to learn richer feature representations. To better learn relevant visual patterns among different SAIs, we treat $L_{LR}$ (or $G_{LR}$) as a collection of $A$ horizontal and $A$ vertical angular sequences, and each sequence includes $A$ consecutive SAIs collected along one direction. Our content (or gradient) Transformer processes these sequences in $L_{LR}$ (or $G_{LR}$)
In particular, in the \( k \)-th block, we treat its input as a set of horizontal (or row-wise) sequences, i.e., \( \mathcal{R} = \{ \mathbf{R}_i \in \mathbb{R}^{A \times 2 \times C \times H \times W} \}_{i=1}^A \), where \( \mathbf{R}_i \) indicates a sequence of convolutional features, corresponding to the \( i \)-th row in the input. The first horizontal SA-LSA layer \( \mathcal{H}_1^H \) aims to explore the dependencies within each sequence independently. Specifically, for each horizontal sequence \( \mathbf{R}_i \), we obtain a non-local representation \( \hat{\mathbf{R}}_i \) as follows:

\[
\hat{\mathbf{R}}_i = \mathcal{H}_1^H(\mathbf{R}_i) \in \mathbb{R}^{A \times C \times H \times W}. \tag{3}
\]

After processing all horizontal sequences in \( \mathcal{R} \), we obtain a horizontal-enhanced content representation for the light field:

\[
\mathbf{R}_k^\text{Cont} = [\hat{\mathbf{R}}_1, \hat{\mathbf{R}}_2, \ldots, \hat{\mathbf{R}}_A] \in \mathbb{R}^{A \times 2 \times C \times H \times W}, \tag{4}
\]

where \([ \ ]\) denotes the concatenation operation.

Next, the second vertical SA-LSA layer \( \mathcal{H}_2^V \) accepts \( \mathbf{R}_k^\text{Cont} \) as input, and explores the long-range relations of vertical (or column-wise) sequences, i.e., \( \mathcal{C} = \{ \mathbf{T}_i \in \mathbb{R}^{A \times 1 \times C \times H \times W} \}_{i=1}^A \), where \( \mathbf{T}_i \) indicates sequences of convolutional features, corresponding to the \( i \)-th column in \( \mathbf{R}_k^\text{Cont} \). Similarly, the vertical sequences in \( \mathcal{C} \) are transformed via \( \mathcal{H}_2^V \) to produce a vertical-enhanced content representation \( \mathbf{T}_k^\text{Cont} \):

\[
\mathbf{T}_k^\text{Cont} = [\hat{T}_1, \hat{T}_2, \ldots, \hat{T}_A] \in \mathbb{R}^{A \times 2 \times C \times H \times W}, \tag{5}
\]

where \( \hat{T}_i = \mathcal{H}_2^V(\mathbf{T}_i) \) is the non-local representation of \( \mathbf{T}_i \). Our content Transformer uses \( K \) spatial-angular attention blocks to aggregate informative contextual knowledge among SAIs, eventually yielding a more comprehensive content representation \( \mathbf{T}_k^\text{Cont} \).

Our gradient Transformer \( \mathbf{T}_k^\text{Grad} \) performs in a similar way as the content Transformer. It accepts \( \mathbf{F}_k^\text{Grad} \) as its input, and utilize \( K \) spatial-angular attention blocks to deliver a gradient representation \( \mathbf{T}_k^\text{Grad} \).

**Cross-Attention Fusion Transformer.** While the content and gradient Transformers process each modality separately, we design a fusion Transformer to aggregate together their representations. To obtain more comprehensive unimodal representations, we obtain the inputs of fusion Transformer.
by combining all intermediate features in the content or gradient Transformers:

\[ \begin{align*}
H^{\text{Cont}} & = [F^{\text{Cont}}, T_1^{\text{Cont}}, \cdots, T_K^{\text{Cont}}] \in \mathbb{R}^{A \times A \times (K+1) \times C \times H \times W}, \\
H^{\text{Grad}} & = [F^{\text{Grad}}, T_1^{\text{Grad}}, \cdots, T_K^{\text{Grad}}] \in \mathbb{R}^{A \times A \times (K+1) \times C \times H \times W}.
\end{align*} \tag{6} \]

Different from \( T^{\text{Cont}} \) and \( T^{\text{Grad}} \) which capture non-local dependencies of tokens within the same sequence, our fusion Transformer aims to explore the relations between tokens of two sequences. In particular, for the \( i \)-th horizontal (or vertical) sequences in \( H^{\text{Cont}} \) and \( H^{\text{Grad}} \), i.e., \( U_i \in \mathbb{R}^{A \times (K+1) \times C \times H \times W} \) and \( V_i \in \mathbb{R}^{A \times (K+1) \times C \times H \times W} \), the \( T^{\text{Fuse}} \) achieves detail-preserved representation \( Z_i \) as:

\[ Z_i = T^{\text{Fuse}}(U_i, V_i) \in \mathbb{R}^{A \times A \times (K+1) \times C \times H \times W}. \tag{7} \]

The \( T^{\text{Fuse}} \) performs cross-attention between its inputs, with query generated from \( U_i \), key and value from \( V_i \), to gather high-frequency information as a compensation of the content representation. We concatenate the outputs \( \{Z_i\}_i \) together to obtain the fusion output \( Z \in \mathbb{R}^{A \times A \times (K+1) \times C \times H \times W} \). Note that our fusion Transformer only includes one spatial-angular attention block, and we see minor performance improvement when adding additional blocks.

**SAI Reconstruction.** Finally, we leverage a reconstruction module \( \mathcal{D}^{\text{Reco}} \) over \( Z \) to obtain a high-resolution LF \( L^{\text{HR}} \):

\[ L^{\text{HR}} = \mathcal{D}^{\text{Reco}}(Z) \in \mathbb{R}^{A \times A \times H \times oW}. \tag{8} \]

Here, \( \mathcal{D}^{\text{Reco}} \) is separately applied to each SAI.

**Remark.** Our DPT employs Transformers to enrich convolutional features \( F^{\text{Cont}} \) and \( F^{\text{Grad}} \) (Eq. 2) into richer representations \( H^{\text{Cont}} \) and \( H^{\text{Grad}} \), respectively, which are further aggregated together via a fusion Transformer. In this manner, our network is able to collect informative non-local contexts within each SAI and across different SAIs, allowing for higher-quality reconstruction.

In addition, our Transformers process each sequence (or sequence fusion) independently rather than directly process all sequences together. This enables our model to meet hardware resource constraints, and more importantly, avoid redundant interactions among irrelevant sub-aperture images.

In Transformer architectures (e.g., ViT (Dosovitskiy et al. 2020)), fully-connected self-attention is extensively employed to explore non-local interactions among input tokens. However, the vanilla self-attention layer neglects local spatial context within each input token, which is crucial for image reconstruction. To remedy this limitation, we introduce a spatial-angular locally-enhanced self-attention layer to offer our Transformer a strong capability in modeling locality.

**Spatial-Angular Locally-Enhanced Self-Attention.** Inspired by VSR (Cao et al. 2021), our SA-LSA layer includes three sequential operations: spatial-angular convolutional tokenization, spatial-angular self-attention, as well as spatial-angular convolutional de-tokenization. Its structure is illustrated in Fig. 2 (a).

**Spatial-Angular Convolutional Tokenization.** ViT-like architectures (Dosovitskiy et al. 2020) leverage a linear projection layer to achieve input tokens at a very early stage. In contrast, our proposed convolutional tokenization mechanism obtains the tokens in the self-attention layer, yielding a multi-stage hierarchy like CNNs. This allows our Transformer to capture local contexts from low-level features to high-level semantic representations.

In particular, denote \( F \in \mathbb{R}^{A \times C \times H \times W} \) as the spatial-angular representation of an angular sequence, where \( A \) is the length of the sequence, \( C \) denotes feature dimension, and \( H \times W \) is the spatial dimension. The convolutional tokenization module generates query \( Q \), key \( K \) and value \( V \) at each self-attention layer as follows:

\[ Q = \mathcal{U}(f_Q(F)), \quad K = \mathcal{U}(f_K(F)), \quad V = \mathcal{U}(f_V(F)). \tag{9} \]

It produces the inputs of a self-attention layer via two steps. First, the input \( F \) is fed into three independent convolutional layers (i.e., \( f_Q, f_K \) and \( f_V \)). We implement each layer with kernel size \( 1 \times 1 \).

Next, a function \( \mathcal{U} \) is designed to obtain spatial-angular tokens. In particular, it extracts a collection of overlapping patches with size \( H_p \times W_p \). Thus, we are able to obtain a sequence \( X \in \mathbb{R}^{n \times d} \) of \( n \) tokens, and each token with a feature dimension \( d = C \times H_p \times W_p \).

**Spatial-Angular Self-Attention.** The fully-connected self-attention layer is applied on \( X \) to explore the non-local spatial-angular relations among the tokens as follows:

\[ X' = \text{Attention}(Q, K, V) = \text{softmax}(QK^T)V, \tag{10} \]

where \( \text{Attention} \) denotes a standard self-attention layer as in (Vaswani et al. 2017; Dosovitskiy et al. 2020).

**Spatial-Angular Convolutional De-Tokenization.** To enable the application of convolutional tokenization in each self-attention layer, we further de-tokenize the attended feature sequence \( X' \) to fold the patches into a large feature map \( \hat{F} \) with the same dimension as \( F \):

\[ \hat{F} = f_p(P(X')) + F \in \mathbb{R}^{A \times C \times H \times W}, \tag{11} \]
where \(\mathcal{P}\) denotes the de-tokenization operation, and \(f_{\mathcal{P}}\) is a convolutional layer as Eq. 9. A residual layer is also used to avoid the loss of important features. Here, \(\hat{F}\) well encodes the local (in the spatial dimension) and global (in both spatial and angular dimensions) context of the input angular sequence, which could be expected to help produce better reconstruction results.

Note that in the fusion Transformer, we improve SA-LSA into a cross-attention SA-LSA layer to support the computation of cross-attention between the two modalities. Fig. 2 shows the detailed structure.

**Detailed Network Architecture**

**Convolutional Feature Extraction.** The convolutional modules \(\mathcal{E}^{\text{Cont}}\) and \(\mathcal{E}^{\text{Grad}}\) (Eq. 2) share a similar network structure. Following (Wang et al. 2021c), each of them consists of two residual blocks and two residual atrous spatial pyramid pooling blocks organized in an interwoven manner, following by an angular alignment module to align the convolutional features of different views.

**Transformer Networks.** In the content and gradient Transformers, we set the number of attention blocks \(K\) to 2 by default. For \(\mathcal{T}^{\text{Fuse}}\), Eq. 9 is slightly modified to enable the exploration of cross-attention by generating query from the content representation, while key and value from the gradient representation.

**Reconstruction Module.** The reconstruction module \(\mathcal{D}^{\text{Reco}}\) (Eq. 8) is implemented as a cascaded of five information multi-distillation blocks, following with an upsampling layer, which consists of a pixel shuffle operation combined with four 2×2 convolution layers, to generate the super-resolved SAIs (Wang et al. 2021c).

**Experiment**

**Datasets and Evaluation Metrics**

We conduct extensive experiments on five popular LFSR benchmarks, i.e., EPFL (Rerabek and Ebrahimi 2016), HCInew (Honauer et al. 2016), HCIold (Wanner, Meister, and Goldluecke 2013), INRIA (Le Pendu, Jiang, and Guillemot 2018), and STFgantry (Vaish and Adams 2008). All of the light field images from the above benchmarks have a 9×9 angular resolution (i.e., \(U = V = A = 9\)). PSNR and SSIM are chosen as the evaluation metrics. For a testing dataset with \(T\) scenes, we first obtain the metric values of \(U \times V\) sub-aperture images, and average the summation results of \(T \times U \times V\) to obtain the final metric score.

**Implementation Details**

Following (Wang et al. 2021c), we convert the light field images from the RGB space to the YCbCr space. Our model only super-resolves Y channel images, and uses the bicubic interpolation to super-resolve Cb and Cr channels images, respectively. The gradient maps are extracted from Y channel of each SAI along the spatial dimension with the function provided by (Ma et al. 2020). We perform ×2 and ×4 SR with 5×5 angular resolution on all benchmarks. In the training stage, the 64×64 patches are cropped from each sub-aperture image, and we apply the bicubic interpolation to generate ×2 and ×4 patches. We use random horizontal rotation, vertical rotation and 90° rotation to augment the training data. Spatial and angular resolution are processed simultaneously for preserving the LF structure. The \(\ell_1\) loss is used to optimize our network. We use the Adam optimizer to train our network, with a batch size of 8. The initial learning rate is set to \(2 \times 10^{-4}\) and it will be halved every 15 epochs. We train the network for 75 epochs in total. All experiments are carried out on a single Tesla V100 GPU card.

**Comparisons with State-of-the-Art**

To evaluate the effectiveness of our method, we compare our DPT with the other state-of-the-art LFSR methods (i.e., LFBM5D (Alain and Smolic 2018), GB (Rossi and Frossard 2018), resLF (Zhang, Lin, and Sheng 2019), LF-SSR (Yeung et al. 2018), LF-InterNet (Wang et al. 2020), and LF-DFNet (Wang et al. 2021c)) and single image super-resolution methods (i.e., VDSR (Kim, Lee, and Lee 2016), EDSR (Lim et al. 2017), and RCAN (Zhang et al. 2018)) on five LFSR benchmarks. Following (Wang et al. 2021c), we treat Bicubic as the baseline method.

**Quantitative Results.** The quantitative performance comparisons are reported in Table 1. As seen, DPT achieves a promising performance compared to other methods. Single image super-resolution methods (i.e., VDSR (Kim, Lee, and Lee 2016), EDSR (Lim et al. 2017), and RCAN (Zhang et al. 2018)) super-resolve each view separately. DPT outperforms all of them, which benefits from the complementary information encoding of different views with our proposed Transformers.

Compared with the convolution-based LFSR methods, e.g., resLF (Zhang, Lin, and Sheng 2019), LFSSR (Yeung et al. 2018), LF-InterNet (Wang et al. 2020), our DPT outperforms all of them for ×4 SR. The reason lies in that convolutional network architectures lack strong capabilities to model global relations among different views. In contrast, our DPT can establish the global relations among different SAIs with the Transformer for efficient spatial-angular representations.

Specifically, compared with the current leading approach LF-DFNet (Wang et al. 2021c), DPT obtains superior results on EPFL (Rerabek and Ebrahimi 2016), HCIold (Wanner, Meister, and Goldluecke 2013) and INRIA (Le Pendu, Jiang, and Guillemot 2018) for ×4 SR. That is because, LF-DFNet only models the short-range dependencies between each side-view SAI and the center-view SAI, while our DPT explores long-range relations among all SAIs.

**Qualitative Results.** Fig. 3 presents the visual comparisons of different methods for ×4 SR. As seen, LF-InterNet (Wang et al. 2020) and LF-DFNet (Wang et al. 2021c) produce distorted results, especially for the structures of the telescope stand in Bicycle scene and the flowerbed edge in Sculpture scene. In contrast, our DPT produces better results, with all above regions being well preserved.

**Computational Analysis.** We compare our model with other LFSR methods in terms of parameters, FLOPs, and reconstruction accuracy. The detailed comparisons are reported in Table 2. Following (Wang et al. 2021c), we also use the input light field image with the size 5×5×32×32 to calculate the
Table 1: Performance comparison of different methods for \( \times 2 \) and \( \times 4 \) SR. The best results are marked as bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scale</th>
<th>EPFL</th>
<th>HCInew</th>
<th>HCOld</th>
<th>INRIA</th>
<th>STFgantry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>( \times 2 )</td>
<td>29.50 / 0.9350</td>
<td>31.69 / 0.9335</td>
<td>37.46 / 0.9776</td>
<td>31.10 / 0.9563</td>
<td>30.82 / 0.9473</td>
</tr>
<tr>
<td>VDSR (Kim, Lee, and Lee 2016)</td>
<td>( \times 2 )</td>
<td>32.50 / 0.9599</td>
<td>34.37 / 0.9563</td>
<td>40.61 / 0.9867</td>
<td>34.43 / 0.9742</td>
<td>35.54 / 0.9790</td>
</tr>
<tr>
<td>EDSR (Lim et al. 2017)</td>
<td>( \times 2 )</td>
<td>33.09 / 0.9631</td>
<td>34.83 / 0.9594</td>
<td>41.01 / 0.9875</td>
<td>34.97 / 0.9765</td>
<td>36.29 / 0.9819</td>
</tr>
<tr>
<td>RCAN (Zhang et al. 2018)</td>
<td>( \times 2 )</td>
<td>33.16 / 0.9625</td>
<td>34.98 / 0.9672</td>
<td>41.05 / 0.9875</td>
<td>35.01 / 0.9769</td>
<td>36.33 / 0.9825</td>
</tr>
<tr>
<td>LFBMSD (Alain and Smolic 2018)</td>
<td>( \times 2 )</td>
<td>31.15 / 0.9545</td>
<td>33.72 / 0.9548</td>
<td>39.62 / 0.9854</td>
<td>32.85 / 0.9695</td>
<td>33.55 / 0.9718</td>
</tr>
<tr>
<td>GB (Rossi and Frossard 2018)</td>
<td>( \times 2 )</td>
<td>31.22 / 0.9591</td>
<td>35.25 / 0.9692</td>
<td>40.21 / 0.9879</td>
<td>32.76 / 0.9724</td>
<td>35.44 / 0.9835</td>
</tr>
<tr>
<td>resLF (Zhang, Lin, and Sheng 2019)</td>
<td>( \times 2 )</td>
<td>32.75 / 0.9672</td>
<td>36.07 / 0.9715</td>
<td>42.61 / 0.9922</td>
<td>34.57 / 0.9784</td>
<td>36.89 / 0.9873</td>
</tr>
<tr>
<td>LF-SSR (Yeung et al. 2018)</td>
<td>( \times 2 )</td>
<td>33.69 / 0.9748</td>
<td>36.86 / 0.9753</td>
<td>43.75 / 0.9939</td>
<td>35.27 / 0.9834</td>
<td>36.91 / 0.9935</td>
</tr>
<tr>
<td>LF-InterNet (Wang et al. 2020)</td>
<td>( \times 2 )</td>
<td>34.34 / 0.9766</td>
<td>37.35 / 0.9770</td>
<td>44.23 / 0.9943</td>
<td>36.36 / 0.9841</td>
<td>39.52 / 0.9928</td>
</tr>
<tr>
<td>LF-DNNet (Wang et al. 2021c)</td>
<td>( \times 2 )</td>
<td>34.44 / 0.9753</td>
<td>37.35 / 0.9770</td>
<td>44.23 / 0.9943</td>
<td>36.36 / 0.9841</td>
<td>39.52 / 0.9928</td>
</tr>
<tr>
<td>DPT (Ours)</td>
<td>( \times 2 )</td>
<td>34.34 / 0.9753</td>
<td>37.35 / 0.9770</td>
<td>44.23 / 0.9943</td>
<td>36.36 / 0.9841</td>
<td>39.52 / 0.9928</td>
</tr>
</tbody>
</table>

Table 2: Detailed comparisons of LFSR methods in terms of number of parameters, FLOPs, and reconstruction accuracy on the EPFL dataset for \( \times 2 \) and \( \times 4 \) SR.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scale</th>
<th># Param (M)</th>
<th>FLOPs (G)</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>resLF</td>
<td>( \times 2 )</td>
<td>6.35 / 37.06</td>
<td>32.75 / 0.9672</td>
<td>38.72 / 0.9916</td>
<td>31.23 / 0.9196</td>
</tr>
<tr>
<td>LFSSR</td>
<td>( \times 4 )</td>
<td>1.61 / 128.44</td>
<td>28.27 / 0.9080</td>
<td>37.32 / 0.9718</td>
<td>30.83 / 0.9503</td>
</tr>
<tr>
<td>LF-InterNet</td>
<td>( \times 4 )</td>
<td>6.79 / 38.70</td>
<td>37.48 / 0.9759</td>
<td>37.32 / 0.9718</td>
<td>30.83 / 0.9503</td>
</tr>
<tr>
<td>LF-DNNet</td>
<td>( \times 4 )</td>
<td>4.80 / 47.46</td>
<td>34.86 / 0.9761</td>
<td>37.32 / 0.9718</td>
<td>30.83 / 0.9503</td>
</tr>
<tr>
<td>DPT (Ours)</td>
<td>( \times 4 )</td>
<td>3.94 / 57.22</td>
<td>34.44 / 0.9766</td>
<td>37.32 / 0.9718</td>
<td>30.83 / 0.9503</td>
</tr>
</tbody>
</table>

FLOPs. As can be seen, DPT achieves the best reconstruct accuracy compared with other LFSR methods. Specifically, compared to LF-DNNet (Wang et al. 2021c), DPT has fewer parameters, while boosts the PSNR by 0.16dB. This result demonstrates the superiority of DPT not only in reconstruction accuracy, but also the lightweight architecture.

Ablation Study

To gain more insights into our model, we conduct a set of ablative experiments on the EPFL dataset for \( \times 4 \) SR. The results are reported in Table 3. We show in the 1st row the performance of baseline model (i.e., LF-DNNet (Wang et al. 2021c)), and the 7th row the results of our full model.

Content Transformer. We first investigate the effect of the content Transformer by constructing a network which is implemented with the convolution feature extraction module \( \xi_{\text{Cont}} \), content Transformer \( T_{\text{Cont}} \), and the image reconstruction module \( D_{\text{Reco}} \). Its results are given in the 2nd row. We can see that the content Transformer itself can lead to a similar performance as the baseline model, however, our content Transformer has fewer parameters. This confirms that global relations among different views brought by the content Transformer are beneficial to LFSR.

Gradient Transformer. Furthermore, we combine the gradient Transformer into the content Transformer, with the results being shown in the 3rd row. As we can see, by introducing the gradient Transformer, the performance of the content Transformer (the 2nd row) improves by 0.09dB in terms of PSNR and 0.0011 improvement in terms of SSIM, respectively. Moreover, the models with dual transformers (the 4th, the 6th and the 7th rows) outperform the model with only a content Transformer (the 2nd row), which further confirms the effectiveness of gradient Transformer. Finally, we replace the content Transformer and gradient Transformer of DPT with residual blocks while maintaining the network parameters almost unchanged. Its results are given in the 5th row. The dual branches in DPT are equivalent to two convolutional neural networks for feature extraction. As reported in Table 3, DPT has a 0.21 dB PSNR drop, demonstrating the effectiveness of the proposed Transformers.

Impact of Fusion Mechanism. To explore the effect of fusion mechanism for content and gradient feature aggregation, we construct two model variants, which generates the detailed-preserved representation \( Z_{i} \) in Eq. 7 by the element-wise summation (the 3rd row) and the Transformer for a single view non-local dependencies exploration (the 4th row), respectively. As can be observed, our cross-attention fusion Transformer (the 7th row) brings a promising performance improvement over the results of 3rd row and the 4th row in PSNR and SSIM, which is attributed to its effectiveness for the complementary non-local information exploration of the content and gradient features.

Efficacy of SA-LSA. We study the effect of the SA-LSA layers in DPT (the 6th row). As seen, the model with SA-LSA layers (the 7th row) obtains a better
Figure 3: Visual comparisons of Bicycle scene from HCInew and Sculpture scene from INRIA for \times 4 SR. We see that our approach achieves compelling reconstruction results, with the details well preserved.

Table 3: Ablation study of DPT on the EPFL dataset for \times 4 SR.

<table>
<thead>
<tr>
<th>#</th>
<th>Content Transformer</th>
<th>Gradient Transformer</th>
<th>Fusion Mechanism</th>
<th># Param (M)</th>
<th>PSNR / SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Transformer (image)</td>
<td>Transformer (sequence)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>Sum</td>
<td>3.99</td>
<td>28.77 / 0.9165</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>2.62</td>
<td>28.77 / 0.9142</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.72</td>
<td>28.86 / 0.9153</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.75</td>
<td>28.81 / 0.9149</td>
</tr>
<tr>
<td>5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.83</td>
<td>28.72 / 0.9139</td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.91</td>
<td>28.89 / 0.9157</td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.78</td>
<td>28.93 / 0.9167</td>
</tr>
</tbody>
</table>

Table 4: Performance and model size comparisons with different numbers of spatial-angular attention block on the EPFL dataset for \times 4 SR.

<table>
<thead>
<tr>
<th>K</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>28.69</td>
<td>28.93</td>
<td>28.77</td>
</tr>
<tr>
<td># Param (M)</td>
<td>2.89</td>
<td>3.78</td>
<td>5.00</td>
</tr>
</tbody>
</table>

This work proposes a Detail Preserving Transformer (DPT), as the first application of a Transformer-based network for LFSR. Instead of leveraging a vanilla fully-connected self-attention layer, we develop a spatial-angular locally-enhanced self-attention layer (SA-LSA) to promote non-local spatial-angular dependencies of each sub-aperture image sequence. Based on SA-LSA, we leverage a content Transformer and a gradient Transformer to learn spatial-angular content and gradient representations, respectively. The comprehensive spatial-angular representations are further processed by a cross-attention fusion Transformer to aggregate the output of the two Transformers, from which a high-resolution light field is reconstructed. We compare the proposed network with other state-of-the-art methods over five commonly-used benchmarks, and the experimental results demonstrate that it achieves favorable performance against other competitors.

**Conclusion**

This work proposes a Detail Preserving Transformer (DPT), as the first application of a Transformer-based network for LFSR. Instead of leveraging a vanilla fully-connected self-attention layer, we develop a spatial-angular locally-enhanced self-attention layer (SA-LSA) to promote non-local spatial-angular dependencies of each sub-aperture image sequence. Based on SA-LSA, we leverage a content Transformer and a gradient Transformer to learn spatial-angular content and gradient representations, respectively. The comprehensive spatial-angular representations are further processed by a cross-attention fusion Transformer to aggregate the output of the two Transformers, from which a high-resolution light field is reconstructed. We compare the proposed network with other state-of-the-art methods over five commonly-used benchmarks, and the experimental results demonstrate that it achieves favorable performance against other competitors.
**Acknowledgments**

This work is supported by the National Key Research and Development Plan (No.2017YFC0112001), and by China Central Television (JG2018-0247).

**References**


