Vision, Modeling, and Visualization (2022)
J. Bender, M. Botsch, and D. Keim (Eds.)

Local Attention Guided Joint Depth Upsampling

Arijit Mallick[1], Andreas Engelhardt[1], Raphael Braun[1] ,Hendrik PA Lensch[1]

[1]University of Tübingen, Department of Computer Graphics, Germany

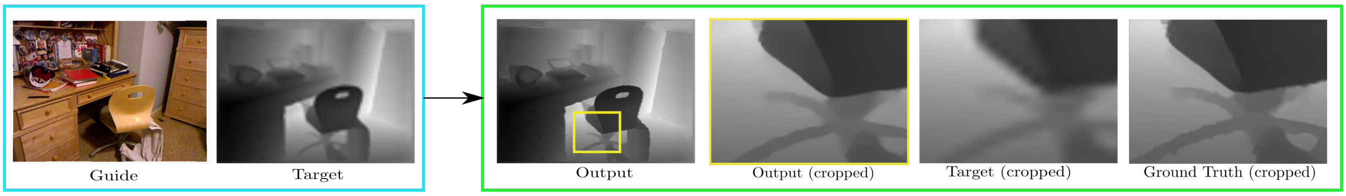


Figure 1: *Our proposed image local attention guided joint depth upsampling network takes the high resolution guide image and the low resolution bicubic upsampled target image as input. Note the upsampling enhancement in the marked patch generated by our network compared to the 8x upsampled input next to it as well as the corresponding ground truth patch.*

Abstract

Image super resolution is a classical computer vision problem. A branch of super resolution tasks deals with guided depth super resolution as objective. Here, the goal is to accurately upsample a given low resolution depth map with the help of features aggregated from the high resolution color image of that particular scene. Recently, the development of transformers has improved performance for general image processing tasks credited to self-attention. Unlike previous methods for guided joint depth upsampling which rely mostly on CNNs, we efficiently compute self-attention with the help of local image attention which avoids the quadratic growth typically found in self-attention layers. Our work combines CNNs and transformers to analyze the two input modalities and employs a cross-modal fusion network in order to predict both a weighted per-pixel filter kernel and a residual for the depth estimation. To further enhance the final output, we integrate a differentiable and a trainable deep guided filtering network which provides an additional depth prior. An ablation study and empirical trials demonstrate the importance of each proposed module. Our method shows comparable as well as state-of-the-art performance on the guided depth upsampling task.

CCS Concepts

• ***Computing methodologies*** → ***Computer vision;*** *Image representations;* ***Reconstruction;***

1. Introduction

Image super resolution (SR) is a classical computer vision problem. Given low resolution input the algorithm tries to compute a corresponding high resolution image. Recent advancements in smartphone photography to satellite imagery employ super resolution for improved image quality and visual clarity. Until now, these methods have been computationally expensive and generally produce low resolution depth maps whose clarity, however, can be increased with RGB image guided depth super resolution. In this paper, we deal with the classic joint depth super resolution (SR) problem. Given a low resolution depth map (target) and a corresponding high resolution RGB image (guide), our task is to compute the corresponding high resolution depth map. Classical depth super resolution methods usually rely on a spatial filtering technique [KCLU07] where the input is upsampled by filtering the local neighborhood with weights directly based on the corresponding patch in the guide image. One of the downsides of this kind of method is that it can be time and memory consuming for very high resolution images. Additionally, it can miss homogeneous background information.

Recent developments in machine learning for computer vision applications have also paved the way for guided depth super resolution methods. In general, these applications try to infer the filter kernel weights for each target pixel with the help of a guide RGB input image to perform an adaptive, spatially-varying convolution on the target image. Inspired by the classical joint depth upsampling task, spatial filter weights for the joint bilateral filter have been replaced by a learnable variant for the multi-view stereo task [YG20] in the past. We take inspiration from this application

DOI: 10.2312/vmv.20221197

and try to infer the neighborhood pixel weights based on an additional local image attention block to extract detailed neighborhood features. Local image attention [YYF*20] has made it possible to cheaply extract expressive image features for this super-resolution tasks.

Our architecture in Figure 2 combines two main ideas. First, we generate an enhanced target input by deep guided filtering networks [WZZH18] and in parallel estimate per-pixel adaptive filter-kernels for upsampling the target images by fusing the features of both the guide and the target. Second, we estimate corresponding residuals which are added onto the guided filtered results to refine the depth-map. Both approaches rely on features which are first extracted separately from the guide and target images then merged for each task employing local attention. The last module of the network incorporates the original low resolution bicubic upsampled target image, guided filtered (GF) target depth and deep guided filtered (DGF) target depth which subsequently proposes the final output from a trainable, weighted per-pixel prediction module.

Features are extracted by a U-Net followed by a self-attention-based transformer encoder to extract local neighborhood information for improved edge-aware guidance. A deep merge network (Mergenet) performs efficient cross-modal fusion of local neighborhood features. By combining the RGB and the depth domain we produce a representation which includes the high-frequency detail from the guide image as well as the coarse depth information from the target image. We use those results twice: As input for the filter-kernels estimation and as input for constructing the residuals from the GF as well as the DGF target map, both of which are a function of the RGB guide image and original bicubic upsampled target image.

Our filter pathway can be interpreted as a generalized adaptive filter with trainable pixel similarity measure. We demonstrate the validity and importance of each module in our ablation study. Our contributions are as follows:

- Local attention and merge block for fusing spatial information from both the guide RGB image and the target depth to provide better super resolution guidance
- Performance comparable to state-of-the-art methods and superior performance in some cases

2. Related Work

Classical methods

The classical joint depth super resolution literature can be divided into filter-based methods and optimization-based approaches. In filter-based approaches, texture and edge features are extracted from the given guide RGB image to inform handcrafted filters that try to estimate the weights for spatially-varying filter masks that are convolved with the lower resolution target image. Joint bilateral upsampling [KCLU07] extends the single image bilateral filter [TM98]) to steer the filter with a guide image. The bilateral weights are obtained by converting the local guide RGB image pixel values to bilateral weights which are then applied cross-modal to the low resolution input. Guided filters [HST13; WZZH18] provide a similar idea of considering a filtered output factor from the guidance image. Aforesaid methods are based on filter kernels where strong local guide features are utilised to enhance a low resolution depth map. The upsampling task has also been addressed as a global energy minimization problem, such as the Markov random field based technique in [DT06]. Non-local means filtering with extended regularization for additional edge weighting has further improved joint depth upsampling [PKT*11]. These methods all employ a regularization term which guides the target towards a structurally similar texture of the high resolution guide image. The fast bilateral solver combines these simple filtering methods and approaches this problem as a domain-specific optimization algorithm [BP16]. Additionally, in [HCP18] static-dynamic filter combinations have shown significant improvements on the joint upsampling task with the help of better structural prior extraction.

Learning-based methods

Contrary to classical techniques which do not rely on supervision, data-driven learning approaches are becoming significantly popular because of their generalisation capability. Early learning-based methods utilised a dictionary in order to express structural similarity within paired guide and target images. Kwon et al. [KTL15] utilize a sparse representation learning of dictionaries on the geometric correlation between high-quality mesh data, ground truth target and guide images. In [YWHM10], a sparse representation of the target map is obtained, and corresponding coefficients are used to predict a high resolution depth output. Lately, CNN-based techniques have shown significant improvements on the task of joint depth super resolution. Multi-scale guidance networks with an encoder-decoder architecture [HLT16] got rid of depth boundary artifacts. Moreover, in [LHAY19], salient structures that are consistent in both guidance and target images are selectively leveraged. The deep primal-dual network [RFRB16] with iterative optimisation has shown better noise removal along with good super-resolution results. Apart from these direct encoder-decoder approaches, The Deformable Kernel Network [KPH20] learns a sparse and spatially-variant kernel which stretches a kernel non-linearly along the given pixel neighborhood. The method in turn extracts a residual offset from the combined image features. Apart from showing better performance, a faster extension was also shown with almost similar metrics [KPH20]. Su et al. [SJS*19] learn to predict the filter weights of a spatially-varying kernel as a function of the local pixel features. A cross-task interaction module is introduced in [SYL*21] to realize bilateral cross-modality knowledge transfer to solve uncertainty depth estimation guided super resolution. In [HZL*21], high-frequency components decomposed from the RGB image subsequently guide the super resolution task. Apart from fully CNN-powered architectures, also densely connected networks have been proposed. [LDWS19] employ a MLP for pixel to pixel mapping of the guide information to the target. Similarly, Tang et al. [TCZ21] utilise a deep implicit neural representation based technique. It is essentially an MLP which efficiently extracts latent codes from the input and appends it to the coordinates, eventually providing a depth correction residual. They achieve state-of-the-art results on noisy joint depth super resolution tasks. Orthogonal work like [dLBD*22] directly optimizes an explicit affinity graph to regularize the reconstruction. Overall, learning-based guided joint upsampling methods usually lever-

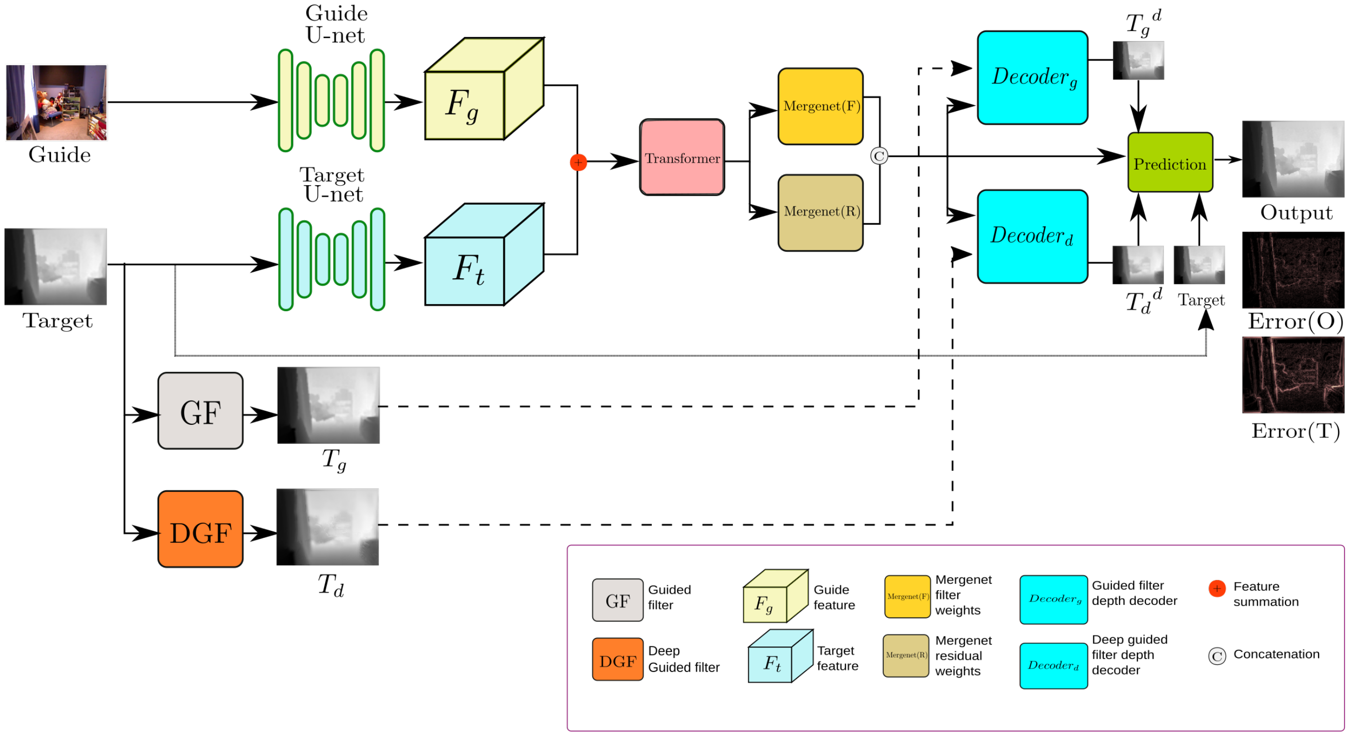


Figure 2: *Attention Guided Upsampling: Extracted features of the guide and the target are fused by a local transformer to predict filter weights as well as an additional residual. This Information is used in the decoder block (shared weights) to predict an upsampled version of T_g and T_d separately. A final prediction layer combines all target predictions. Error(0) and Error(T) visualize the difference to the original and the target input image.*

age monocular depth-like datasets [SHKF12; LRL14; SP07; HS07; HZL*21]. Learnt joint bilateral upsampling has been integrated into the multi-view stereo task [YG20] where the bilateral weights are selected as a function of the given reference image for sparse-to-dense depth approximation which significantly reduces the computation effort and provides a faster reconstruction. Contrary to the existing networks, we contribute additional refinement to the low resolution guided depth map inputs with the help of transformer encoded attention weights. Additionally, our residual network contributes stronger edge aware features.

Transformers and local attention

Transformers [VSP*17] have become a widely used architecture, especially in Natural Language Processing [DCLT19; BMR*20]. Transformers primarily operate with the concept of self-attention, which explores the relation between all tokens in a sequence to capture contextual information. The base transformer encoder models have been successfully applied to low-level computer vision tasks such as classification [DBK*20]. Recently, the Texture Transformer Network for image super resolution [YYF*20] uses low resolution and reference RGB images as queries and keys in a transformer. They essentially transfer the high resolution texture to a low resolution image for a super resolution task. In the context of guided depth super resolution self-attention has just started to be explored as a part of larger architectures [XCW*21; YCZT22; AC22]. The Discrete Cosine Transform module in [ZZX*22] employs an edge attention mechanism to highlight the contours which provides useful information for guided upsampling. As basic self-attention has quadratic complexity in the number of tokens Longformer [BPC20b] introduces a number of different sampling approaches that improve the efficiency of attention evaluations. In particular, the local sliding window attention mechanism scales linearly with the sequence length, allowing it to process even very large token sets. This idea can also be found in [ZZX*22] where grouped convolutions are used to compute attention maps to weight edge information. In our scenario, we apply local sliding-window attention to a 2d patch around a pixel. Local attention provides a weighting of the spatially combined guide and target feature tensors which helps in extracting rich contextual information. A separate merge network further enhances the correlation between them, leveraging both the power of CNNs and transformers for an efficient depth residual computation.

3. Method

The goal in joint depth upsampling is to use a high resolution *guide image* for adding missing detail in an aligned low resolution *target image*. Our network solves this task in four major steps: Guided depth proposals, feature extraction, cross-modality merging and final guided filtering. In addition to the low resolution target image, we obtain a guided filtered target and a deep guided filtered tar-

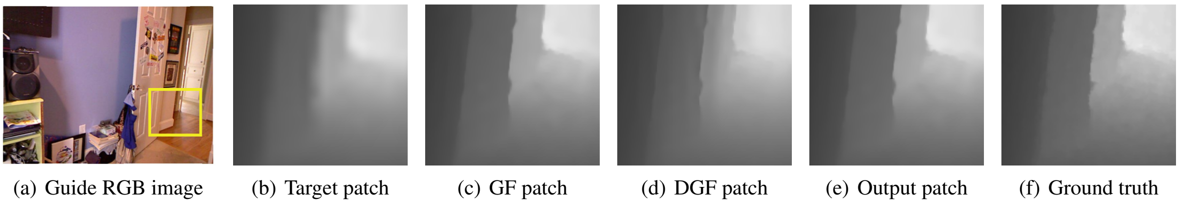


Figure 3: *We demonstrate the contribution of the guided filter blocks. While the target patch is clearly affected by the upsampling artefacts, guided filter map post residual refinement clearly shows remarkable improvements wrt. feature sharpness. Deep guided filtering contributes additional edge aware features. Finally, the weighted pixel prediction module provides additional improvements in the output patch and when compared to ground truth, the overall upsampling artefacts are hardly noticeable.*

get [WZZH18]. During feature extraction the guide image and target image are processed independently to generate feature vectors for each pixel. In the merging stage the features from the guide image are combined with the target features to produce the inputs to the filter section. The GF target and the DGF target are filtered with adaptive kernels and augmented with a residual estimate to the detailed *output* depth map. The final output is a pixel-wise weighted combination of the original target, filtered GF target and the filtered DGF target. In this section we describe our architecture from Figure 2 in detail.

Depth guided filters

Direct bicubic upsampling of low resolution images produces significant artefacts as it does not consider the spatial context. In order to provide better input, we utilise two simple guided filter modules. The first block is a differentiable Guided Filter (GF) layer which takes the low resolution target image I_t and high resolution image I_g to generate a high resolution proposal $T_g = GF(I_t, I_g)$ by a linear transformation [HST13]. The second block consists of a Deep Guided Filter network which integrates the previous guided transformation layers into CNNs and generates corresponding guidance maps $T_d = DGF(I_t, I_g)$. See [WZZH18] for further details on the gradient propagation through guided filtering and convolutional guided filtering layers.

Feature extraction

Upsampling an image is inherently a local operation, however in order to fill in local details it can help to consider the global context, such as reoccurring patterns or regularities in the occurrence of depth discontinuities. We use U-Nets [RFB15] f_g and f_t to extract primary features $F_g = f_g(I_g)$ and $F_t = f_t(I_g)$, with separate weights for the guide image I_g and target image I_t. I_t has been upsampled with a bicubic filter to the same resolution. Those features are based on the local neighborhood of each pixel in different scales. It is to be noted that we do not extract features for T_d and T_g, as they are already jointly encoded with vital guide and target image information and F_g, F_t have sufficient information for further operations in the rest of the architecture.

Next, spatial self-attention compares and relates each stacked pixel feature (F_g, F_t) against its neighbors to better judge its relative importance and to localize important information for the final task of edge-aware upsampling. The self-attention is only computed locally over a sliding window similar to 2D-convolutions [CGRS19; BPC20a; RR20] but with content dependant filter weights. The query, key and value for the attention mechanism are extracted using a linear transformation across the channel dimension which is implemented as 1×1 convolutions. With a quadratic window of side-length p (here $p = 5$) the memory requirement of local self-attention is limited to $O(n * p^2) = O(n)$, where n is the number of pixels in the input. Since the patch-size is a constraint for varying and higher resolution cases, a combined feature map would provide richer edge-aware pixel neighborhood information during attention computation in the following stage. Hence, we use a transformer encoder [VSP*17] $\mathscr{T}t$ block built with the aforementioned local attention to enrich our spatially combined target and guide features with detailed local information. The final features $\mathscr{A} = \mathscr{T}(F_g + F_t)$ are computed by applying the transformer to the combined U-Net-feature-maps.

Mergenet

During feature extraction there is no cross-talk between the information extracted from the guide image I_g and target image I_t. Even though the transformer encoder block enhances the combined features for depth guidance, only self-attention is not sufficient. The mergenet is responsible to not only combine both modalities, but provides further enhanced guidance cues. The Mergenet consists of 8 2D convolution layers with ReLU blocks as activations. It consists of two separate blocks (F and R), both working on the same input, that produce the weights W_F and W_R needed for the *Filter* and *Residual* steps to create the final depth prediction.

Depth decoder

The decoder module combines the result of a *Filter* module $\mathscr{F}$ with a separately computed depth *Residual* $\mathscr{R}$. The adaptive filter module converts W_F into a per-pixel filter kernel which is convolved with the guided target images T_g and T_d. As the adaptive filter can only produce a weighted average of already existing depth values the residual module estimates a depth-correction from W_R and W_F. Here, W_R can potentially infuse some additional information from the guide image features estimated in the Mergenet. We apply the same operation with shared weights to T_g and T_d separately.

$$T_{g,d}{}^{d} = \mathscr{F}_{g,d}(T_{g,d}, W_F) + \mathscr{R}_{g,d}(W_F, W_R) \quad (1)$$

Depth filter module $\mathscr{F}$ The joint bilateral upsampler [KCLU07] has been employed as a classical solution for guided depth upsampling. This method uses a range filter and a spatial filter for predicting the filtered depth output. We take inspiration from its learnt counterpart in FastMVSNET [YG20] who implicitly try to encode the spatial information with the help of a simple CNN. We utilise a learned version to filter the low resolution target with kernels constructed from W_F, rather than directly using image features. This can be viewed as a generalized adaptive upsampler with an estimated kernel for every pixel coordinate.

To predict the adaptive per-pixel filter mask, we reduce the dimensionality of W_F with the help of a simple 1×1 convolutional network f_{reduce} and utilize it to convolve the target image and obtain $W_{reduce} = softmax(f_{reduce}(W_F))$.

Note that W_{reduce} has k^2 channels, where k is the chosen kernel-size in the filter module. Let $N_k(x)$ be the list of indices in the pixel neighborhood centered at x, then the filter operation can be written as:

$$\mathscr{F}_{g,d}(T_{g,d}, W_F)[x] = \sum_{i=1}^{k^2} W_{reduce}[x,i] \cdot T_{g,d}[N_k(x)[i]] \tag{2}$$

Depth residual module $\mathscr{R}$ The depth values produced by the filter are formed by building kernels which are convolved with the low-detail target images. However, the features produced in the Mergenet module already contain the detailed information from the guide image as well as the depth information from the target image. We therefore use W_R and W_F directly to compute an additional residual, which is added to the filter result as indicated in Equation 1. With W_R and W_F having the same spatial and channel dimensions we can combine them in an element-wise multiplication and sum up the channels to produce a one-channel residual map. This module can be interpreted as a simple pixel-wise weighted residual prediction from the filter and residual weights (W_R, W_F). We will take cues from the original aligned feature maps in order to provide proper weights for the different target proposals. We interpret W_F as a confidence score for the residual contribution W_R and hence compute the overall weighted residual as :

$$\mathscr{R}_{g,d}(W_F, W_R) = \sum_{i=1}^{C} \text{softmax}(W_F) \cdot W_R \tag{3}$$

where C is the feature channel dimension of W_R and W_F.

Depth prediction

The final module is a simple pixel-wise weighted depth prediction $(pred)$ module that estimates the final output from the three proposed depth maps $(I_t, {T_g}^d, {T_c}^d)$. We will take cues from the original guide feature map and just computed W_R and W_F in order to provide proper weights for the different target proposals. This block consists of 4 convolution layers which estimate the final weights $W_{pred} = softmax(pred(F_g, W_F, W_R))$. Thus, the final upsampled depth prediction is given as:

$$D_{final} = \sum \text{softmax}(W_{pred}) \cdot (I_t, {T_g}^d, {T_d}^d) \tag{4}$$

Loss function

Given the ground truth high resolution target image D_{gt}, and the network output as D_{final}, we train our network with a L1 loss.

$$Loss = \frac{1}{N_p} \sum_{y=1}^{N_p} |(D_{final} - D_{gt})| \tag{5}$$

Here, N_p is the total number of pixels in the target image.

Table 1: *Ablation study for 8x resolution on NYUv2: Numbers indicate RMSE (lower the better) for the case of 8x bicubic upsampling.*

Method	RMSE
Without transformer	2.80
Without Mergenet	2.95
Without prediction block (mean)	2.79
Without dgf	2.81
Without cdgf	2.73
Without filters	2.84
Without residuals	2.85
Ours	**2.71**

4. Experiments and Results

Datasets and training setup

Our network is trained on NYUv2 [SHKF12] training dataset which consists of 1000 images. We test our trained network on the test split of NYUv2 consisting of 449 images, following the established split protocol of [KPH20]. Additionally, we also test our network on [LRL14] test split and [SP07] test split, following the test convention of [LRL14; KPH20]. It is to be noted that the network is trained separately with 4x, 8x, and 16x downsampling as input following the mentioned conventions. The downsampled target image is upsampled with the help of bicubic upsampling and is used as an input along with the high resolution RGB guide image. We use a NVIDIA RTX3090 to train our network for approximately 14 hours. Keeping in mind the massive self-attention computation cost which involves memory cost proportional to $p \times p$ per pixel, we use an efficient implementation without rearranging keys and values in memory with custom CUDA kernels for the attention computations [Zha19].

Hyperparameters

For the experiments presented in the following sections, our feature channel dimension is set to 128. The NYUv2 training dataset is trained on its full resolution of 480x640 pixels at a batch size of 1. We use 1e-3 as the learning rate for the Adam optimizer. We further decay our learning rate by a factor of 0.2 every 22 epochs. We use a patch size of 5 for the image local attention in all scale scenarios. Our number of heads for the transformer encoder block is set to 1 and the dimension of the feedforward channels is 128. The filter kernel size is set at 7 for all upscale factors. The network is trained end-to-end for 50 epochs.

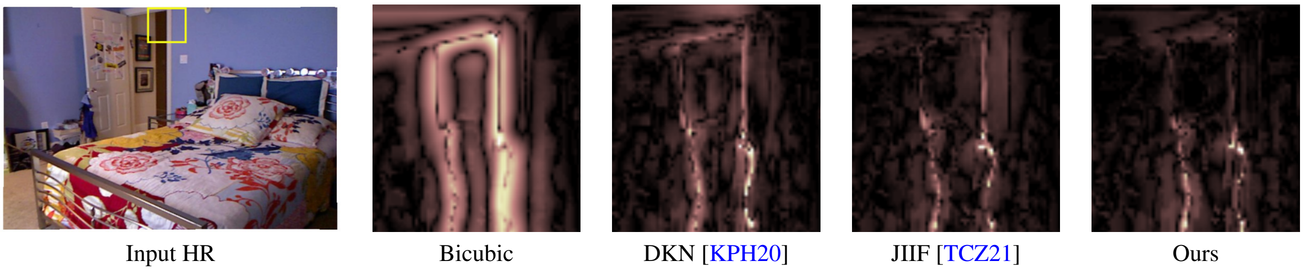
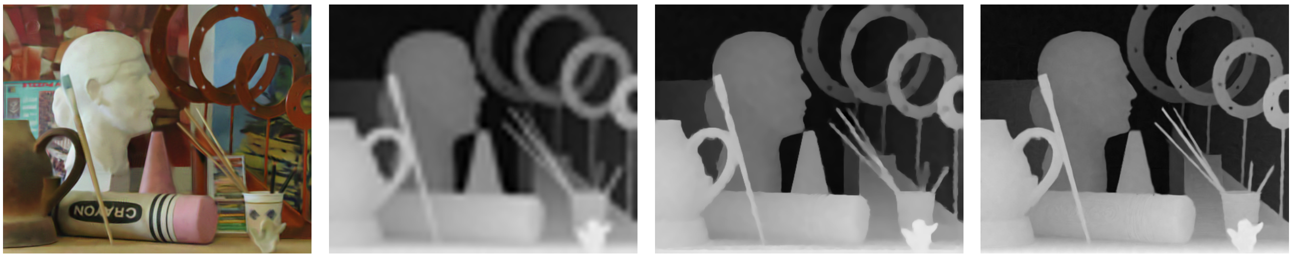


Figure 4: *Qualitative analysis of joint depth upsampling with the help of our network. We demonstrate (5x) upscaled absolute error maps with respect to the ground truth for the patches marked in green in the input HR (High Resolution) images. We compare our network output with DKN [KPH20] and JIIF [TCZ21] on the NYUv2 [SHKF12] test dataset. Brighter regions indicate higher error.*



(a) Guide RGB image I_g (b) Bicubic upsampled target input I_t (c) Network output D_{final} (d) Ground truth depth map D_{gt}

Figure 5: *8x bicubic upsampled example from the Middlebury dataset. Notice the bicubic upsampling artifacts in the target input image(b). Figure (c) shows visual improvement over the target input. Ground truth (d) is given for reference.*

Quantitative results

We compare our method with other learnable joint upsampling algorithms on multiple test datasets (NYUv2, Lu, Middlebury) as shown in Table 2. Our algorithm's performance is comparable and near superior to state-of-the art methods with regard to Root Mean Square Error (RMSE). Our network achieves state of the art on the 4x upsampling task of the NYUv2 dataset, Middlebury dataset; and on 8x upsampling task of NYUv2 and Lu dataset. It demonstrates near state of the art performance for 16x upsampling task with respect to the leading methods which proves our method's generalisation capabilities. Although we outperform JIIF [TCZ21] on multiple test sets for 4x, 8x upsampling, performance drops slightly behind JIIF in 16x upsampling test scenarios. With the increased upsampling factor (16x) the neighborhood context of a fixed-sized local attention encoder is decreased relative to the target image size. We identify this as the main cause for the limited performance. Nevertheless, our network performs effectively well in comparison to leading methods within a reasonable training period, as we are runner up to JIIF [TCZ21] for Middlebury and Lu test datasets.

Qualitative results

We provide a qualitative comparison of the visual clarity of our results with examples from different datasets. In addition, Figure 4 shows a comprehensive comparison with DKN [KPH20]. Additionally, one can also visualize the NYUv2 [SHKF12] test image input along with the 8x upsampled bicubic target input. Overall, one can notice that our network provides a sharper depth output compared to the naive upsampling as well as to the advanced DKN [KPH20] approach. For a more comprehensive insight on the generalisation, we have also provided the results on the Middlebury test set [SP07] in Figure 5. Compared to the degraded (8x bicubic upsampled) input and the corresponding ground truth, the network output is able to preserve sharp details and only introduced very few interpolation artifacts. Visually, the network recovers a substantial amount of depth data in all settings and displays low absolute error along edges of image structures.

Ablation study

To investigate the importance of the individual components in our network, we perform an ablation study by removing each of the six primary training modules from our overall architecture. As presented in Table 1, removing the transformer or the depth filter hinders the performance of our network. Additionally, one can also observe that the absence of the residual module significantly deteriorates the performance as the enhanced cross-model transfer from the target embedding during the final depth computation at the end of the pipeline is missing. Additionally, if we do not enhance the transformer fused target and the guide embedding with the help of our proposed Mergenet, the network struggles to transfer the high resolution texture features to the final depth. Additionally, absence of the depth prediction module also highlights the need of careful pixel selection provided by the fused guidance weights. Introducing GF and DGF provides a much needed prediction prior which again infuses the guide RGB features from the beginning and helps the network to predict the final depth from a better target stand-

Table 2: *Quantitative evaluation (lower is better) for different methods. The evaluation is done in accordance with conventional evaluation metric protocols [KPH20; TCZ21]. Here, the RMSE is taken in units of centimeter. Best results are in blue, and second best results are highlighted in pink.*

Method	NYUv2			Middlebury			Lu		
	4x	8x	16x	4x	8x	16x	4x	8x	16x
Bicubic	4.28	7.14	11.58	2.28	3.98	6.37	2.42	4.54	7.38
DMSG([HLT16], from [KPH20])	3.02	5.38	9.17	1.88	3.45	6.28	2.30	4.17	7.22
DJF([LHAY16], from [KPH20])	2.80	5.33	9.43	1.68	3.24	5.62	1.65	3.96	6.75
DJFR([LHAY19], from [KPH20])	2.38	4.94	9.18	1.32	3.19	5.57	1.15	3.57	6.77
PAC([HLT16], from [KPH20])	1.89	3.33	6.78	1.32	2.62	4.58	1.20	2.33	5.19
DKN[KPH20]	1.62	3.26	6.51	1.23	2.12	4.24	0.96	2.16	5.11
FDSR[HZL*21]	1.61	3.18	5.86	1.13	2.08	4.39	1.29	2.19	5.00
CTKT[SYL*21]	1.49	**2.73**	**5.11**	-	-	-	-	-	-
DCTNet[ZZX*22]	1.59	3.16	5.84	1.10	2.05	4.19	**0.88**	1.85	4.39
JIIF[TCZ21]	**1.37**	2.76	**5.27**	**1.09**	**1.82**	**3.31**	**0.85**	**1.73**	**4.16**
Ours	**1.34**	**2.71**	5.39	**1.07**	**1.86**	**3.57**	0.89	**1.73**	**4.25**

point. This underlines the effect and importance of all the proposed modules in our pipeline.

5. Conclusion

We propose a novel architecture to combine the power of CNNs and transformer-based encoders to solve the guided depth upsampling task with efficient local image attention. Our network consists of a local attention block for extracting edge-aware features from both input modalities, followed by merge networks for cross-modal fusion. To predict the final depth map we extend a set of learned adaptive filters by adding a novel depth residual computation. This increases the sharpness of the upsampled depth map.The approach yields state-of-the-art results in smaller upsampling cases and performs well on larger upsampling tasks when compared to leading methods. We tune the local attention patch size for the optimal trade-off between compute time and performance. An ablation study demonstrates how each sub-module of our network architecture plays an important role in understanding, gathering and subsequently merging the image features. In future work, we would like to improve performance over a wider range of upscaling factors, minimizing the effort for retraining. For example, certain parts of the network can be fine-tuned to accommodate for different input scales while the large U-nets stay fixed. Also a training scheme that trains on multiple datasets and upsampling factors at the same time can improve the generality of the model.

Acknowledgement

This work has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC number 2064/1 - project number 390727645 and SFB 1233 - project number 276693517. It was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and Cyber Valley. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Arijit Mallick.

References

[AC22] Ariav, Ido and Cohen, Israel. "Depth Map Super-Resolution via Cascaded Transformers Guidance". *Frontiers in Signal Processing*. 2022 3.

[BMR*20] Brown, Tom B., Mann, Benjamin, Ryder, Nick, et al. "Language Models are Few-Shot Learners". (2020). arXiv: 2005.14165 [cs.CL] 3.

[BP16] Barron, Jonathan T and Poole, Ben. "The Fast Bilateral Solver". *ECCV* (2016) 2.

[BPC20a] Beltagy, Iz, Peters, Matthew E, and Cohan, Arman. "Longformer: The long-document transformer". *arXiv preprint arXiv:2004.05150* (2020) 4.

[BPC20b] Beltagy, Iz, Peters, Matthew E., and Cohan, Arman. *Longformer: The Long-Document Transformer*. 2020. arXiv: 2004.05150 [cs.CL] 3.

[CGRS19] Child, Rewon, Gray, Scott, Radford, Alec, and Sutskever, Ilya. "Generating long sequences with sparse transformers". *arXiv preprint arXiv:1904.10509* (2019) 4.

[DBK*20] Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2020. arXiv: 2010.11929 [cs.CV] 3.

[DCLT19] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Burstein, Jill, Doran, Christy, and Solorio, Thamar. Association for Computational Linguistics, 2019, 4171–4186. DOI: 10.18653/v1/n19-1423 3.

[dLBD*22] De Lutio, Riccardo, Becker, Alexander, D'Aronco, Stefano, et al. "Learning Graph Regularisation for Guided Super-Resolution". en. (2022), 10 2.

[DT06] Diebel, James and Thrun, Sebastian. "An Application of Markov Random Fields to Range Sensing". *Advances in Neural Information Processing Systems*. Ed. by Weiss, Y., Schölkopf, B., and Platt, J. Vol. 18. MIT Press, 2006 2.

[HCP18] Ham, Bumsub, Cho, Minsu, and Ponce, Jean. "Robust Guided Image Filtering Using Nonconvex Potentials". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.1 (2018), 192–207. DOI: 10.1109/TPAMI.2017.2669034 2.

[HLT16] HUI, TAK-WAI, LOY, CHEN CHANGE, and TANG, XIAOOU. "Depth Map Super-Resolution by Deep Multi-Scale Guidance". *Computer Vision – ECCV 2016*. Ed. by LEIBE, BASTIAN, MATAS, JIRI, SEBE, NICU, and WELLING, MAX. Cham: Springer International Publishing, 2016, 353–369. ISBN: 978-3-319-46487-9 2, 7.

[HS07] HIRSCHMULLER, HEIKO and SCHARSTEIN, DANIEL. "Evaluation of Cost Functions for Stereo Matching". *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 2007, 1–8. DOI: 10.1109/CVPR.2007.383248 3.

[HST13] HE, KAIMING, SUN, JIAN, and TANG, XIAOOU. "Guided Image Filtering". *IEEE Trans. Pattern Anal. Mach. Intell.* 35.6 (2013), 1397–1409. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2012.213 2, 4.

[HZL*21] HE, LINGZHI, ZHU, HONGGUANG, LI, FENG, et al. "Towards Fast and Accurate Real-World Depth Super-Resolution: Benchmark Dataset and Baseline". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, 9229–9238 2, 3, 7.

[KCLU07] KOPF, JOHANNES, COHEN, MICHAEL F., LISCHINSKI, DANI, and UYTTENDAELE, MATT. "Joint Bilateral Upsampling". *ACM Trans. Graph.* 26.3 (2007), 96–es. ISSN: 0730-0301. DOI: 10.1145/1276377.1276497 1, 2, 5.

[KPH20] KIM, BEOMJUN, PONCE, JEAN, and HAM, BUMSUB. "Deformable Kernel Networks for Joint Image Filtering". *International Journal of Computer Vision* 129.2 (2020), 579–600. ISSN: 1573-1405. DOI: 10.1007/s11263-020-01386-z. URL: http://dx.doi.org/10.1007/s11263-020-01386-z 2, 5–7.

[KTL15] KWON, HYEOKHYEN, TAI, YU-WING, and LIN, STEPHEN. "Data-driven depth map refinement via multi-scale sparse representation". *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, 159–167. DOI: 10.1109/CVPR.2015.7298611 2.

[LDWS19] LUTIO, RICCARDO DE, D'ARONCO, STEFANO, WEGNER, JAN DIRK, and SCHINDLER, KONRAD. "Guided Super-Resolution As Pixel-to-Pixel Transformation". en. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, 8828–8836. ISBN: 978-1-72814-803-8. DOI: 10.1109/ICCV.2019.00892 2.

[LHAY16] LI, YIJUN, HUANG, JIA-BIN, AHUJA, NARENDRA, and YANG, MING-HSUAN. "Deep Joint Image Filtering". *Computer Vision – ECCV 2016*. Ed. by LEIBE, BASTIAN, MATAS, JIRI, SEBE, NICU, and WELLING, MAX. Cham: Springer International Publishing, 2016, 154–169. ISBN: 978-3-319-46493-0 7.

[LHAY19] LI, YIJUN, HUANG, JIA-BIN, AHUJA, NARENDRA, and YANG, MING-HSUAN. "Joint image filtering with deep convolutional networks". *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2019), 1909–1923 2, 7.

[LRL14] LU, SI, REN, XIAOFENG, and LIU, FENG. "Depth Enhancement via Low-Rank Matrix Completion". *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, 3390–3397. DOI: 10.1109/CVPR.2014.433 3, 5.

[PKT*11] PARK, JAESIK, KIM, HYEONGWOO, TAI, YU-WING, et al. "High quality depth map upsampling for 3D-TOF cameras". *2011 International Conference on Computer Vision* (2011), 1623–1630 2.

[RFB15] RONNEBERGER, OLAF, FISCHER, PHILIPP, and BROX, THOMAS. "U-Net: Convolutional Networks for Biomedical Image Segmentation". *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by NAVAB, NASSIR, HORNEGGER, JOACHIM, WELLS, WILLIAM M., and FRANGI, ALEJANDRO F. Cham: Springer International Publishing, 2015, 234–241. ISBN: 978-3-319-24574-4 4.

[RFRB16] RIEGLER, GERNOT, FERSTL, DAVID, RÜTHER, MATTHIAS, and BISCHOF, HORST. "A Deep Primal-Dual Network for Guided Depth Super-Resolution". *CoRR* abs/1607.08569 (2016). arXiv: 1607.08569 2.

[RR20] RAE, JACK and RAZAVI, ALI. "Do Transformers Need Deep Long-Range Memory?": *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. URL: https://www.aclweb.org/anthology/2020.acl-main.672 4.

[SHKF12] SILBERMAN, NATHAN, HOIEM, DEREK, KOHLI, PUSHMEET, and FERGUS, ROB. "Indoor segmentation and support inference from rgbd images". *European conference on computer vision*. Springer. 2012, 746–760 3, 5, 6.

[SJS*19] SU, HANG, JAMPANI, VARUN, SUN, DEQING, et al. "Pixel-Adaptive Convolutional Neural Networks". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 2.

[SP07] SCHARSTEIN, DANIEL and PAL, CHRIS. "Learning Conditional Random Fields for Stereo". *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 2007, 1–8. DOI: 10.1109/CVPR.2007.383191 3, 5, 6.

[SYL*21] SUN, BAOLI, YE, XINCHEN, LI, BAOPU, et al. "Learning Scene Structure Guidance via Cross-Task Knowledge Transfer for Single Depth Super-Resolution". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, 7792–7801 2, 7.

[TCZ21] TANG, JIAXIANG, CHEN, XIAOKANG, and ZENG, GANG. "Joint Implicit Image Function for Guided Depth Super-Resolution". *Proceedings of the 29th ACM International Conference on Multimedia* (2021). DOI: 10.1145/3474085.3475584 2, 6, 7.

[TM98] TOMASI, C. and MANDUCHI, R. "Bilateral filtering for gray and color images". *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*. 1998, 839–846. DOI: 10.1109/ICCV.1998.710815 2.

[VSP*17] VASWANI, ASHISH, SHAZEER, NOAM, PARMAR, NIKI, et al. "Attention is All you Need". *Advances in Neural Information Processing Systems*. Ed. by GUYON, I., LUXBURG, U. V., BENGIO, S., et al. Vol. 30. Curran Associates, Inc., 2017 3, 4.

[WZZH18] WU, HUIKAI, ZHENG, SHUAI, ZHANG, JUNGE, and HUANG, KAIQI. "Fast End-to-End Trainable Guided Filter". *CVPR*. 2018 2, 4.

[XCW*21] XING, XIAOXIA, CAI, YINGHAO, WANG, YANQING, et al. "Dynamic Guided Network for Monocular Depth Estimation". *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), 5459–5465 3.

[YCZT22] YANG, YUXIANG, CAO, QI, ZHANG, JING, and TAO, DACHENG. "CODON: On Orchestrating Cross-Domain Attentions for Depth Super-Resolution". en. *International Journal of Computer Vision* 130.2 (Feb. 2022), 267–284. ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-021-01545-w. (Visited on 09/06/2022) 3.

[YG20] YU, ZEHAO and GAO, SHENGHUA. "Fast-MVSNet: Sparse-to-Dense Multi-View Stereo With Learned Propagation and Gauss-Newton Refinement". *CVPR*. 2020 1, 3, 5.

[YWHM10] YANG, JIANCHAO, WRIGHT, JOHN, HUANG, THOMAS S., and MA, YI. "Image Super-Resolution Via Sparse Representation". *IEEE Transactions on Image Processing* 19.11 (2010), 2861–2873. DOI: 10.1109/TIP.2010.2050625 2.

[YYF*20] YANG, FUZHI, YANG, HUAN, FU, JIANLONG, et al. "Learning Texture Transformer Network for Image Super-Resolution". *CVPR*. 2020 2, 3.

[Zha19] ZHANG, ZHENDONG. *Image Local Attention: a Better PyTorch Implementation*. 2019 5.

[ZZX*22] ZHAO, ZIXIANG, ZHANG, JIANGSHE, XU, SHUANG, et al. "Discrete Cosine Transform Network for Guided Depth Map Super-Resolution". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, 5697–5707 3, 7.