Supplementary Material for GuideFormer: Transformers for Image Guided Depth Completion

Kyeongha Rho[1*] Jinsung Ha[2*] Youngjung Kim[1†]
[1]Agency for Defense Development (ADD), Daejeon, Korea
[2]LUXROBO, Seoul, Korea
{khrho325, read12300}@add.re.kr, jinsung@luxrobo.com

Abstract

In this supplementary material, we provide a detailed explanation for GuideFormer and additional analysis for the related experiments that have not been presented in the main paper due to space constraints. This includes a detailed explanation of the dual-branch transformer-based encoder-decoder architecture and qualitative analysis of GuideFormer.

1. Model Architecture of GuideFormer

In this section, we first explain multi-head self- and guided-attention mechanisms, then introduce the more detailed architecture of GuideFormer.

1.1. Multi-head Self- and Guided-attention Mechanisms

In the main paper, we present the implementation of the self- and guided-attention mechanism. Here, we explain the multi-head attention mechanism. As introduced in [2], multi-head attention performs attention h times in parallel with different linearly projected queries, keys, and values. By performing multi-head attention, a transformer model can jointly capture various information from different token representation subspaces. Given a input feature $\mathbf{F} \in \mathbb{R}^{N \times C}$, query, key, and value for multi-head self-attention are first linearly projected into $\frac{C}{h}$ dimensions,

$$\mathbf{Q}_i = \mathbf{F}\mathbf{W}_i^q,\ \mathbf{K}_i = \mathbf{F}\mathbf{W}_i^k,\ \mathbf{V}_i = \mathbf{F}\mathbf{W}_i^v\ , \tag{1}$$

where $i \in \{1,\, 2,\, \cdots,\, h\}$ and $\mathbf{W}_i^q, \mathbf{W}_i^k, \mathbf{W}_i^v \in \mathbb{R}^{C \times \frac{C}{h}}$. The self-attention output for each head is computed as follows:

$$\text{S-Attn}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \text{softmax}(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d_h}} + \mathbf{B})\mathbf{V}_i\ , \tag{2}$$

where $\mathbf{B}$ is the relative positional bias [1] and $d_h = \frac{C}{h}$ is the channel dimension of each head. Then the all h outputs are concatenated into a single final output:

$$\begin{aligned} \mathbf{F}_{out} &= \text{Concat}(\mathbf{H}_1, \cdots, \mathbf{H}_h) \\ \text{where} \quad \mathbf{H}_i &= \text{S-Attn}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) \end{aligned} \tag{3}$$

The multi-head guided-attention is similar to the multi-head self-attention, except that the guided-attention takes multi-modal features as inputs. Given an input feature $\mathbf{F}_I \in \mathbb{R}^{N_I \times C_I}$ and a guidance feature $\mathbf{F}_G \in \mathbb{R}^{N_G \times C_G}$, they are projected into query, key, and value separately.

$$\mathbf{Q}_i = \mathbf{F}_I\mathbf{W}_i^q,\ \mathbf{K}_i = \mathbf{F}_G\mathbf{W}_i^k,\ \mathbf{V}_i = \mathbf{F}_G\mathbf{W}_i^v\ , \tag{4}$$

where $\mathbf{W}_i^q \in \mathbb{R}^{C_I \times \frac{C_G}{h}}$, $\mathbf{W}_i^k, \mathbf{W}_i^v \in \mathbb{R}^{C_G \times \frac{C_G}{h}}$. The guided attention output for each head is computed as follows:

$$\text{G-Attn}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \text{softmax}(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d_h}} + \mathbf{B})\mathbf{V}_i \tag{5}$$

Then the final output of the guided-attention is given as

$$\begin{aligned} \mathbf{F}_{out} &= \text{Concat}(\mathbf{H}_1, \cdots, \mathbf{H}_h)\mathbf{W}^o \\ \text{where} \quad \mathbf{H}_i &= \text{G-Attn}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i), \end{aligned} \tag{6}$$

and $\mathbf{W}^o \in \mathbb{R}^{C_G \times C_I}$. In this work we set h to $\frac{C}{32}$ for C in each encoder and decoder stage.

1.2. Two Different Architectures of GuideFormer

Figure 1 shows more detailed illustration of GuideFormer with two guidance architectures (i.e. parallel and sequential). As mentioned in the main paper, GuideFormer consists of three main components: (1) a fully transformer-based encoder-decoder architecture, (2) guided-attention module (GAM), and (3) the depth fusion module. The difference between the two guidance architectures is the inputs of the GAMs. In the parallel guidance architecture, the GAMs take the color and the depth tokens from the encoders of two branches as their inputs. In the sequential guidance architecture, the GAMs take the color tokens from the decoder of the color branch, while they take the depth

Model	Patch emb.	DWC	Upsample	Skip	RMSE
default	res. blocks	O	shuffle	O	**765.38**
(a)	strided conv.	O	shuffle	O	798.46
(b)	res. blocks	X	shuffle	O	768.94
(c)	res. blocks	O	bilinear	O	767.71
(d)	res. blocks	O	shuffle	X	779.85

Table 1. Ablation studies for the transformer components.

tokens from the encoder of the depth branch for their inputs. The tokens from the encoder and the decoder of the color branch contain different contextual information and their compatibility with the depth tokens as well. This leads to the difference in the effect of guidance, as shown in ablation studies in the main paper.

2. More Experimental Results

2.1. Ablation Studies for the Design of Transformer

In this section, we provide the ablation studies for the distinct components of our fully transformer-based enc-dec, including modified patch embedding, DWC, upsample, and skip connection, in the supplementary material (as in Table 1). We found that, similar to CNN-based enc-dec, skip connections are very important for dense prediction by bridging the gap between enc-dec representations. Other components are also useful, and improve the performance of dense depth completion.

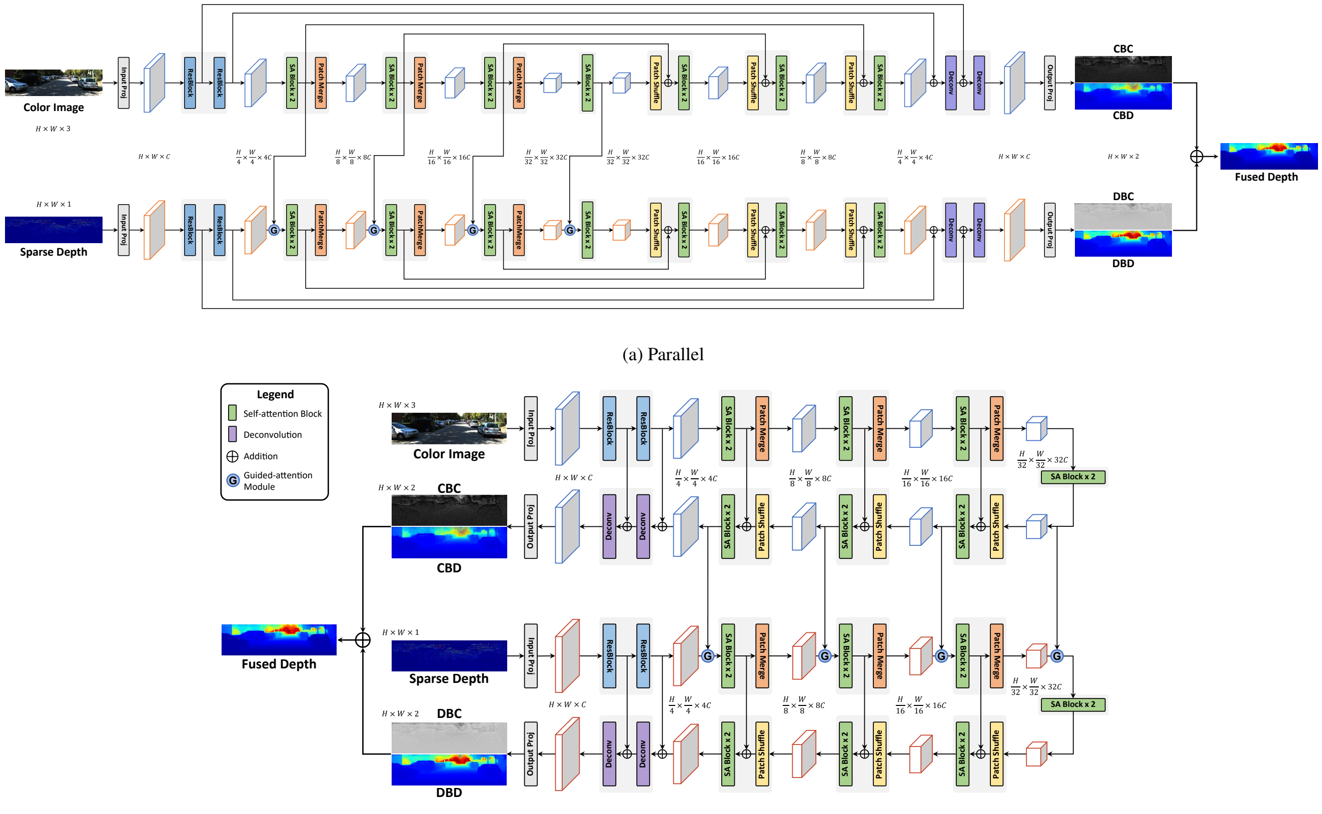
2.2. Qualitative Analysis for GuideFormer

In the main paper, we provide quantitative analysis for ablation studies on GuideFormer: (1) the transformer encoder-decoder architecture, (2) GAM variants, and (3) the guidance architectures. In this section, we present qualitative results for them, as shown in Figure 2, 3, and 4. We perform ablation studies with the validation set. We pick a variety of objects in the images for fair comparison such as bikes, traffic signs, trees, and cars.

References

[1] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *arXiv preprint arXiv:2103.14030*, 2021. 1

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, page 6000–6010, 2017. 1



(a) Parallel

(b) Sequential

Figure 1. Detailed illustration of GuideFormer with two guidance architectures. Here, we express the pre-guide GAM as a representative.

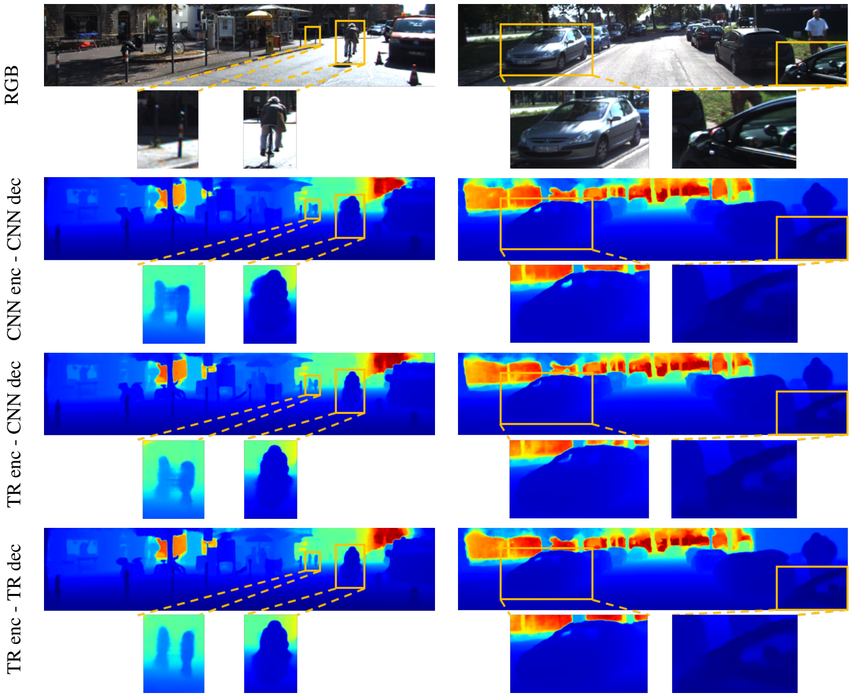


Figure 2. Qualitative results for ablation studies on GuideFormer - CNN vs. Transformer (TR). Concatenation is used for fusing multi-modal information by the sequential architecture in all models.

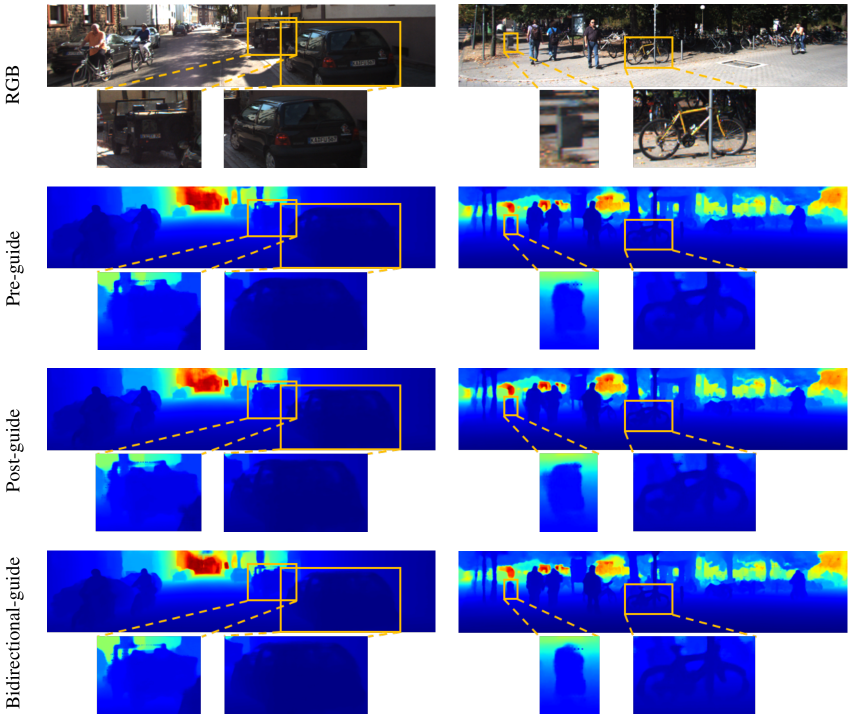


Figure 3. Qualitative results for ablation studies on GuideFormer - GAM variants. All models adopt the parallel guidance architecture for fair comparison.

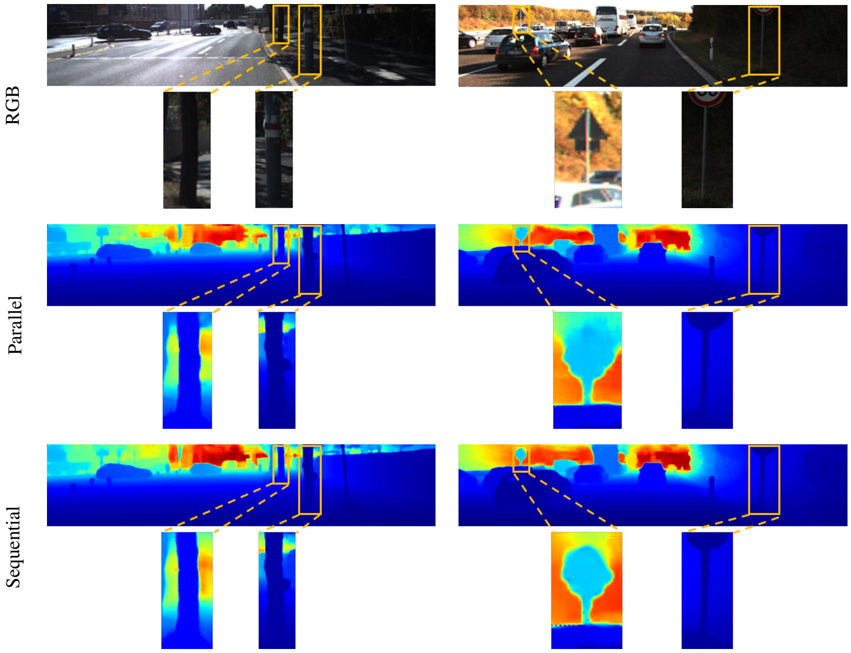


Figure 4. Qualitative results for ablation studies on GuideFormer - the guidance architecture. Pre-guide GAMs are used for comprising the models.