Supplementary Material: Neural Distributed Image Compression with Cross-Attention Feature Alignment

Nitish Mital[1,*], Ezgi Özyılkan[1,†], Ali Garjani[1,‡], and Deniz Gündüz[*]

[*]*Dept. of Electrical and Electronics Engineering, Imperial College London*
[†]*Dept. of Electrical and Computer Engineering, New York University*
[‡]*Section of Mathematics, EPFL*
{*n.mital, d.gunduz*}*@imperial.ac.uk, eo2135@nyu.edu, ali.garjani@epfl.ch*

1 Number of parameters and inference times

Model	Parameter Count (M)	Inference Time (s)
Ours (ATN)	30.23	0.1404
NDIC (w/ Ballé2017 baseline)	16.32	0.0454
DSIN	10.12	0.0414
Ballé2018	10.14	0.0331
Ballé2017	3.94	0.0327

Table 1: Number of parameters and average inference times per test image associated with each considered DNN-based image compression scheme. The inference times are obtained by running the models on a GeForce RTX 2080 GPU with 11 GB RAM. We see that although achieving better rate-distortion performances across various datasets, incorporating additional cross-attention modules within our proposed approach increases the number of parameters as well as the average testing time.

2 CAM parameters

We briefly discuss the parameters used in CAM, which was explained in detail in Section 2.3. We opt for the parameter choices of $C_p = 1$, and $W_p = 8, H_p = 4$ for the first CAM, $W_p = 16, H_p = 8$ for the second CAM and $W_p = 32, H_p = 16$ for the third CAM layers. We choose $D_1 = 256$ for the dimension parameter of the patch embeddings, and $D_2 = 256$ for the dimension parameter of query, key and value vectors. We employ 8 parallel attention heads.

3 Sample pairs of correlated images

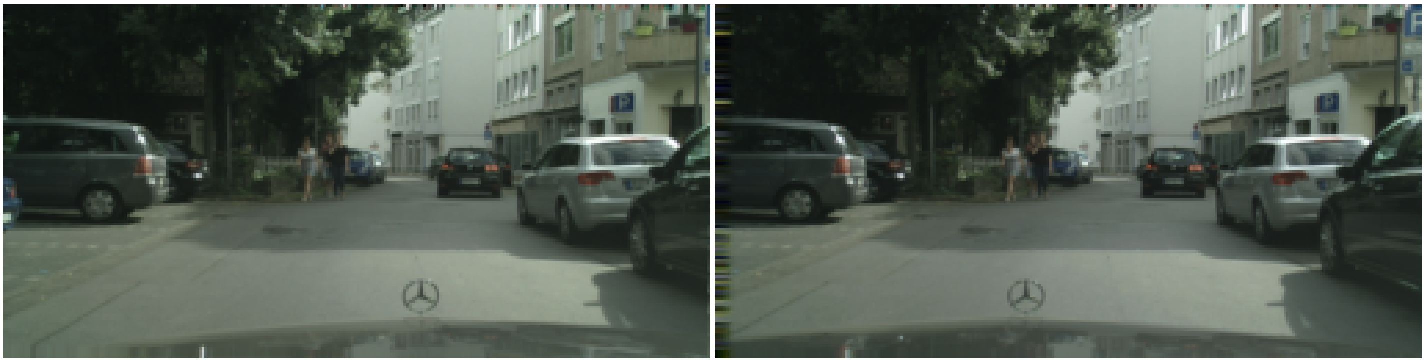


Figure 1: A sample pair of correlated images from the *Cityscape* dataset. Left and right images show the original image and the side information one, respectively.

[1] Contributed equally to this work.

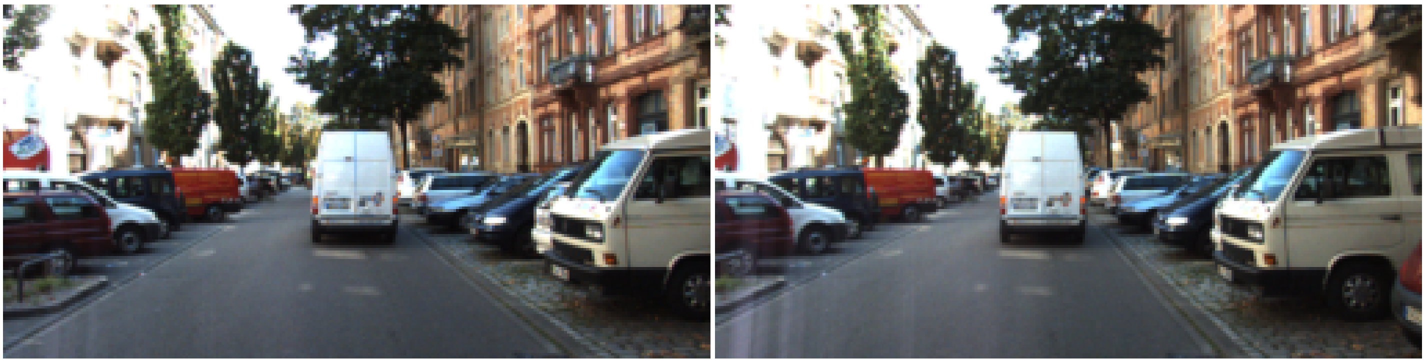


Figure 2: A sample pair of correlated images from the *KITTI Stereo* dataset. Left and right images show the original image and the side information one, respectively.

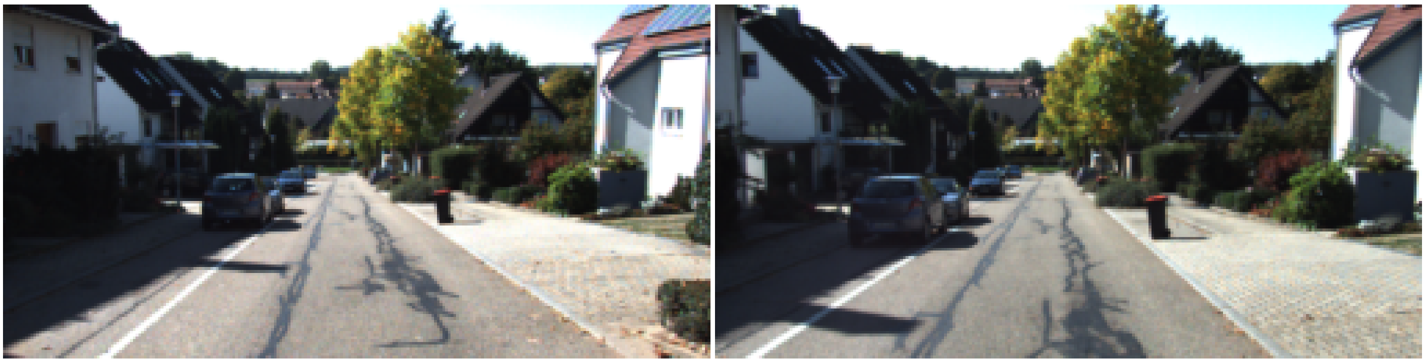


Figure 3: A sample pair of correlated images from the *KITTI General* dataset. Left and right images show the original image and the side information one, respectively. Compared to the two other datasets, we can note that the correlation is less clear, and the correlation relationship is not the same among different pairs of correlated images (see Fig. 4). Hence, it is more challenging for the model to exploit such a correlation while reconstructing the original image.

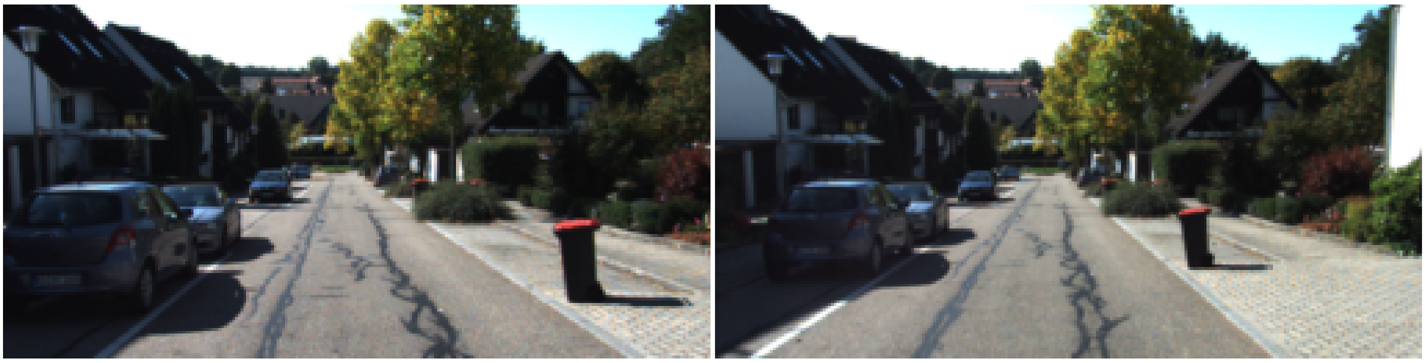
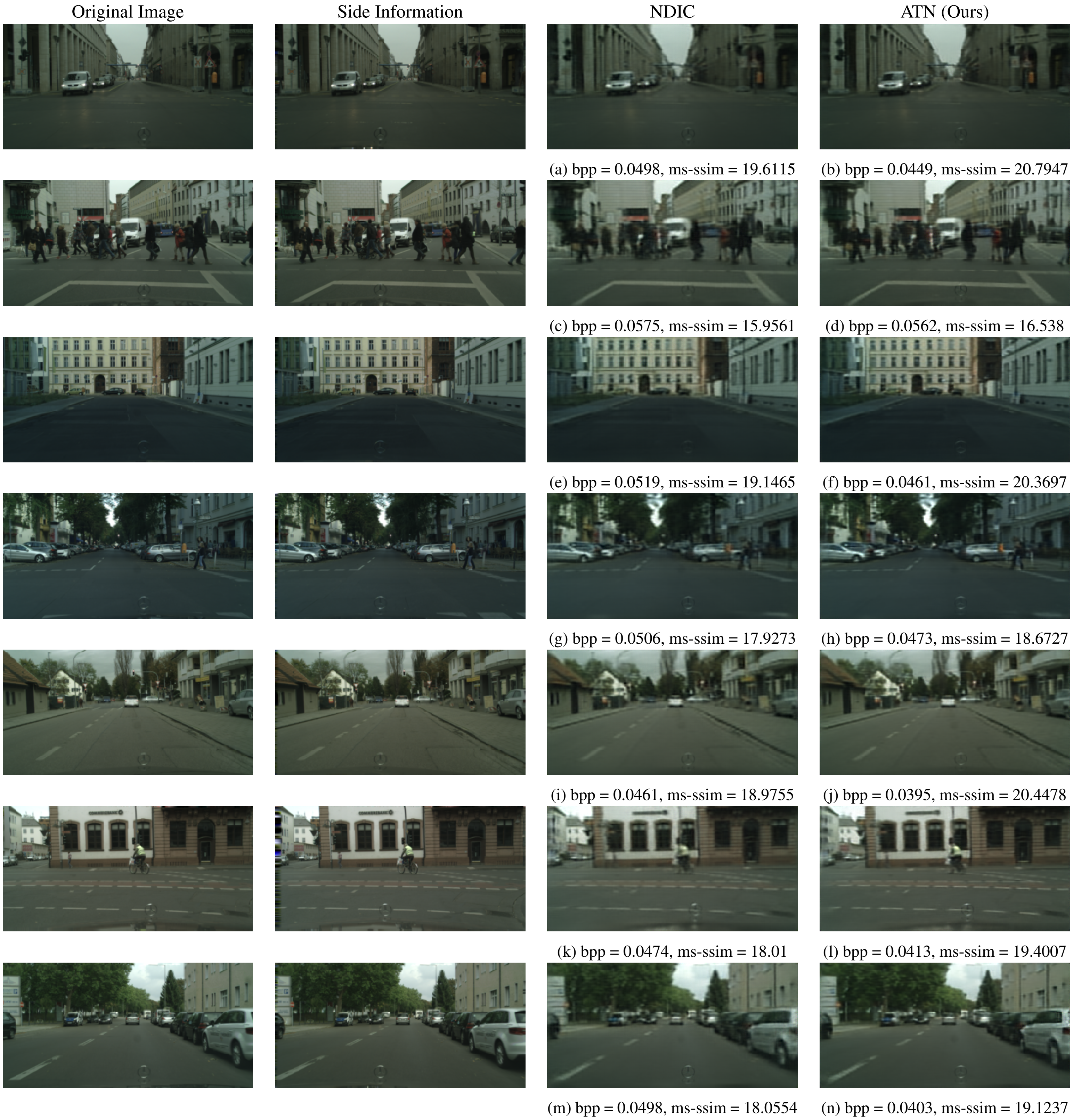


Figure 4: Another sample pair from the *KITTI General* dataset. This sample represents a pair of images that were captured at a different time step from the pair in Fig. 3

4 Additional Visual Proofs

4.1 Cityscape



Original Image

Side Information

NDIC

ATN (Ours)

(a) bpp = 0.0498, ms-ssim = 19.6115

(b) bpp = 0.0449, ms-ssim = 20.7947

(c) bpp = 0.0575, ms-ssim = 15.9561

(d) bpp = 0.0562, ms-ssim = 16.538

(e) bpp = 0.0519, ms-ssim = 19.1465

(f) bpp = 0.0461, ms-ssim = 20.3697

(g) bpp = 0.0506, ms-ssim = 17.9273

(h) bpp = 0.0473, ms-ssim = 18.6727

(i) bpp = 0.0461, ms-ssim = 18.9755

(j) bpp = 0.0395, ms-ssim = 20.4478

(k) bpp = 0.0474, ms-ssim = 18.01

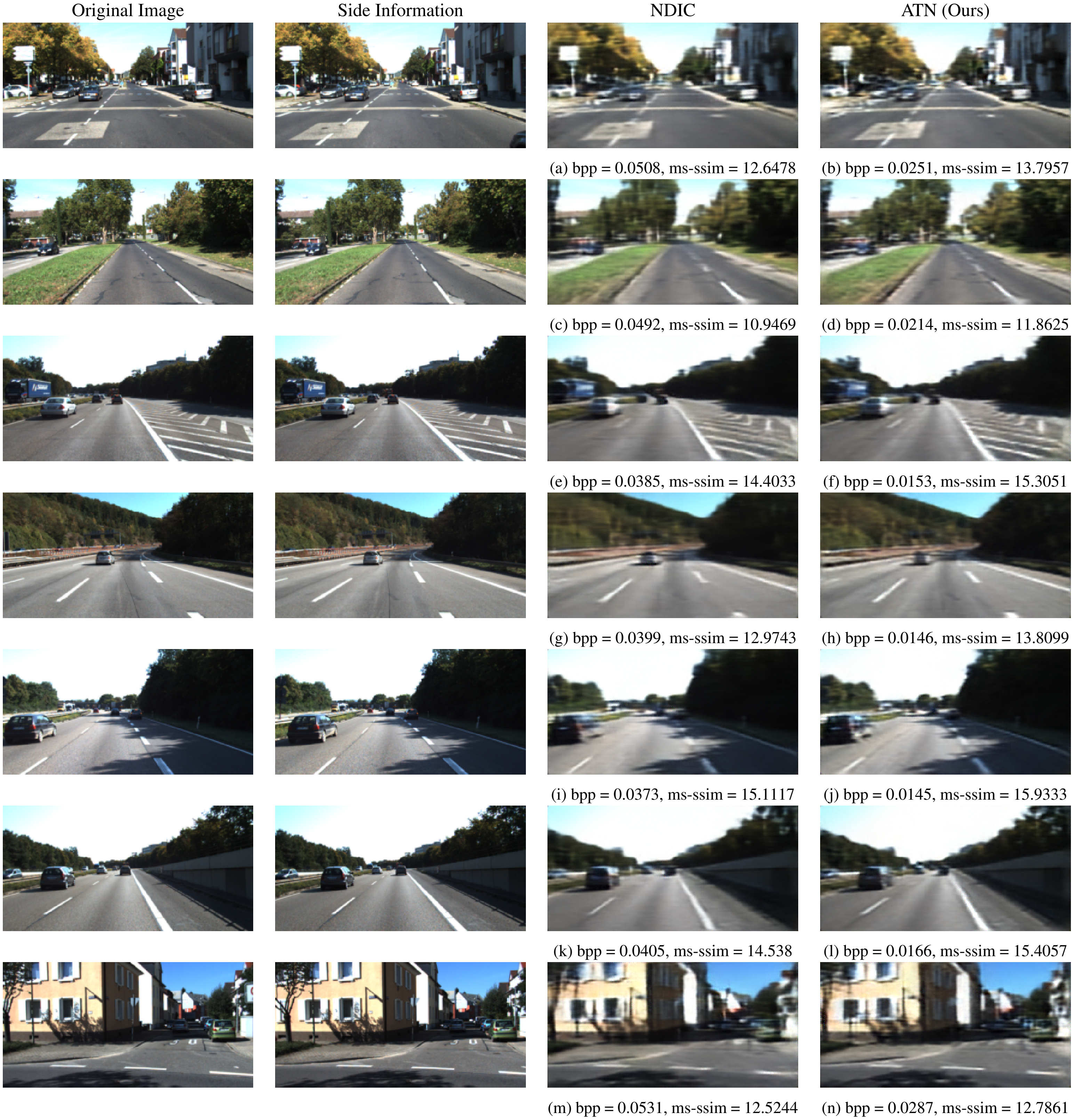
(l) bpp = 0.0413, ms-ssim = 19.4007

(m) bpp = 0.0498, ms-ssim = 18.0554

(n) bpp = 0.0403, ms-ssim = 19.1237

Figure 5: Visual examples comparing NDIC and our model on the Cityscape dataset.

4.2 KITTI Stereo



Original Image

Side Information

NDIC

ATN (Ours)

(a) bpp = 0.0508, ms-ssim = 12.6478

(b) bpp = 0.0251, ms-ssim = 13.7957

(c) bpp = 0.0492, ms-ssim = 10.9469

(d) bpp = 0.0214, ms-ssim = 11.8625

(e) bpp = 0.0385, ms-ssim = 14.4033

(f) bpp = 0.0153, ms-ssim = 15.3051

(g) bpp = 0.0399, ms-ssim = 12.9743

(h) bpp = 0.0146, ms-ssim = 13.8099

(i) bpp = 0.0373, ms-ssim = 15.1117

(j) bpp = 0.0145, ms-ssim = 15.9333

(k) bpp = 0.0405, ms-ssim = 14.538

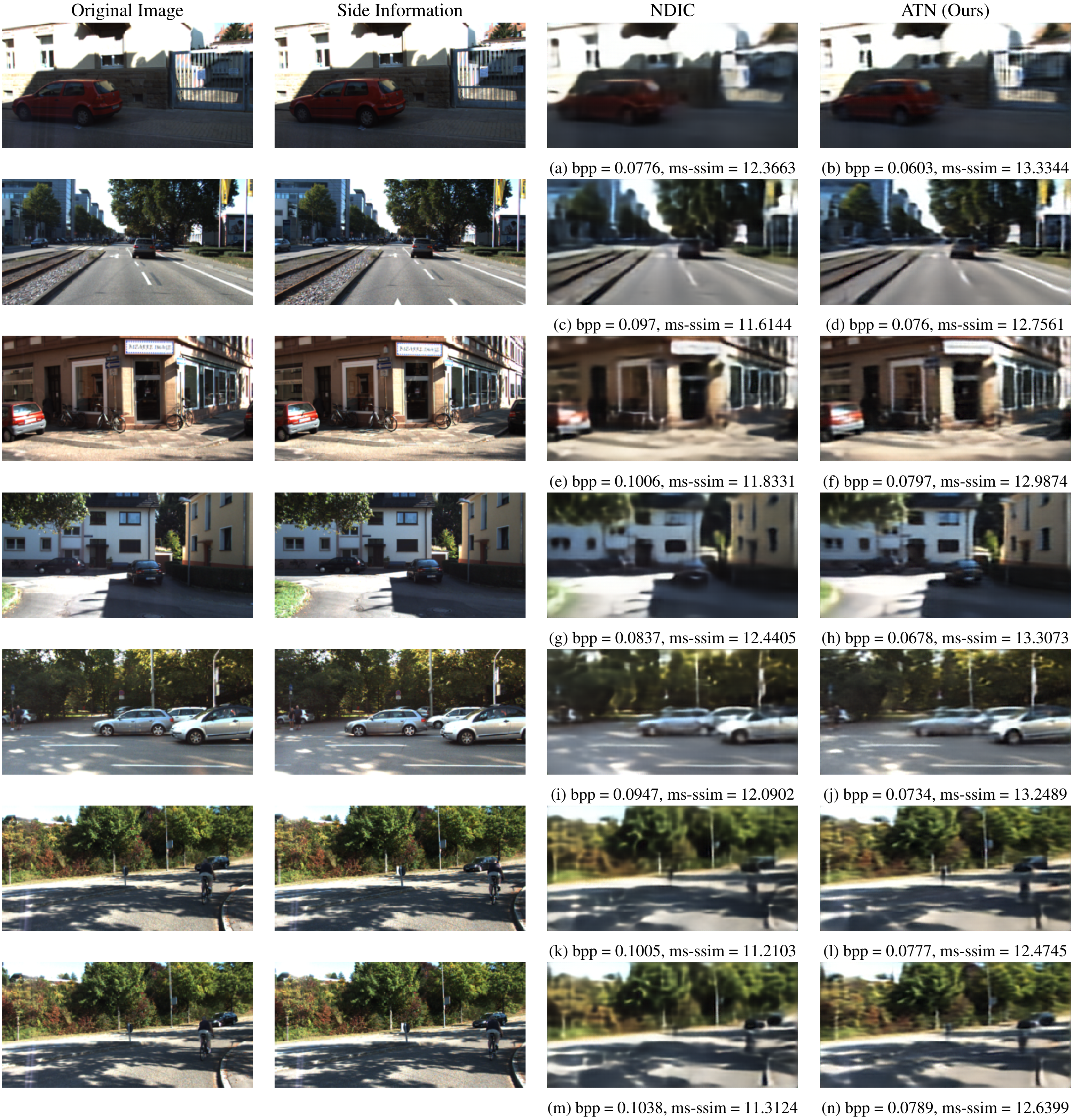
(l) bpp = 0.0166, ms-ssim = 15.4057

(m) bpp = 0.0531, ms-ssim = 12.5244

(n) bpp = 0.0287, ms-ssim = 12.7861

Figure 6: Visual examples comparing NDIC and our model on the KITTI Stereo dataset.

4.3 KITTI General



Original Image

Side Information

NDIC

ATN (Ours)

(a) bpp = 0.0776, ms-ssim = 12.3663

(b) bpp = 0.0603, ms-ssim = 13.3344

(c) bpp = 0.097, ms-ssim = 11.6144

(d) bpp = 0.076, ms-ssim = 12.7561

(e) bpp = 0.1006, ms-ssim = 11.8331

(f) bpp = 0.0797, ms-ssim = 12.9874

(g) bpp = 0.0837, ms-ssim = 12.4405

(h) bpp = 0.0678, ms-ssim = 13.3073

(i) bpp = 0.0947, ms-ssim = 12.0902

(j) bpp = 0.0734, ms-ssim = 13.2489

(k) bpp = 0.1005, ms-ssim = 11.2103

(l) bpp = 0.0777, ms-ssim = 12.4745

(m) bpp = 0.1038, ms-ssim = 11.3124

(n) bpp = 0.0789, ms-ssim = 12.6399

Figure 7: Visual examples comparing NDIC and our model on the KITTI General dataset. Image pairs from two last rows were captured at different time steps