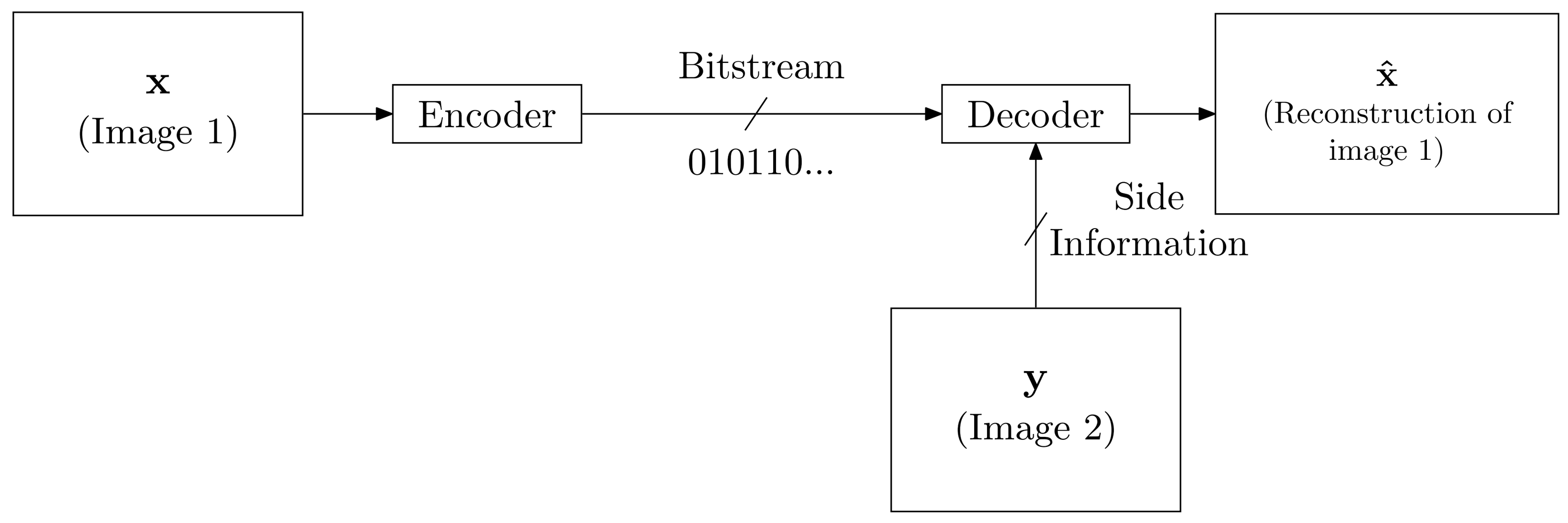
Neural Distributed Image Compression with Cross-Attention Feature Alignment

***Nitish Mital**[*,1]**, Ezgi Ozyilkan**[†,1]**, Ali Garjani**[‡,1]**, Deniz Gunduz**[*]*

[*]*Dept. Of Electrical and Electronics Engineering, Imperial College London*
[†]*Dept. Of Electrical and Computer Engineering, New York University*
[‡]*Section of Mathematics, EPFL*
[1]*Equal contribution*

System Model



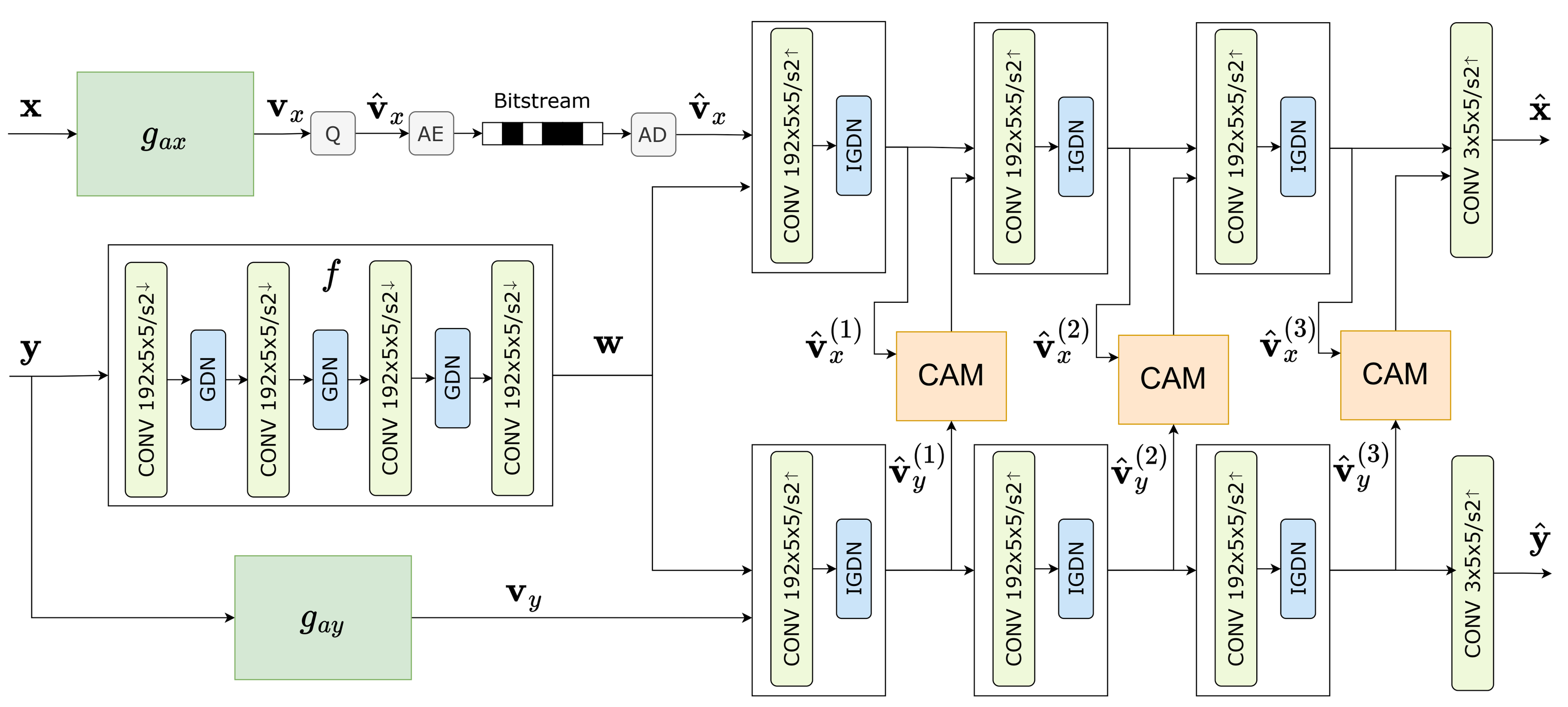
Motivation: Decoder-only side information can provide immense reductions in the transmission rate in lossy compression schemes [1] !

Real-life applications include distributed sensor networks (e.g., autonomous vehicles, multiple static cameras, unmanned aerial vehicles).

Distributed Image Compression

- **DSIN [2] :** Finds corresponding patches to refine the reconstructed image.
- **NDIC [3] :** Extracts ``common information" between correlated images.
- **Our work**, **ATN** : Employs cross-attention modules (CAMs) to align intermediate latent representations.
  - Computes the attention globally, between patches of the latent representations over all channels, similarly to [4].
  - This is similar to *patch-matching* idea in [2], but our method provides a ***differentiable*** alternative to search-based algorithm used in [2].

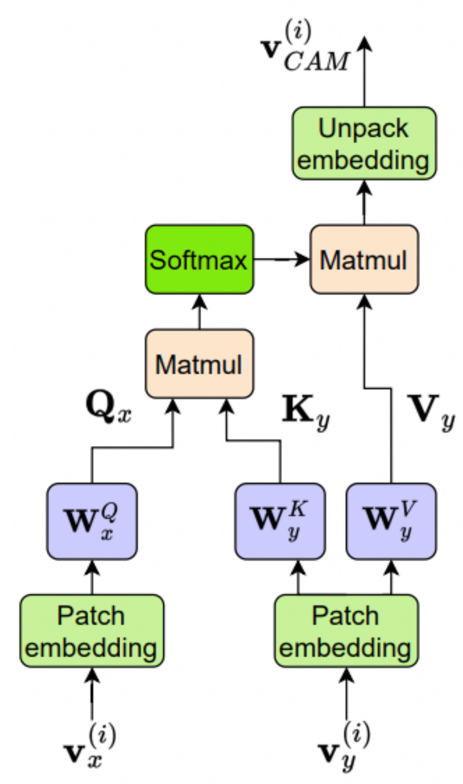
Proposed Architecture



- $\mathbf{w}$ - common features of two images
- $\mathbf{v}_x$, $\mathbf{v}_y$ - local/private features
- Extract $\mathbf{w}$ from $\mathbf{y}$, send only $\mathbf{v}_x$

- Align intermediate latents $\mathbf{v}_x^{(i)}$ and $\mathbf{v}_y^{(i)}$ (in i^{th} layer) using cross-attention module (CAM)
- Generate *query* $\mathbf{Q}_x$ from $\mathbf{v}_x^{(i)}$, *key* $\mathbf{K}_y$ and *value* $\mathbf{V}_y$ from $\mathbf{v}_y^{(i)}$ (all learnable weight matrices!)

$$\mathscr{L} = (R_x + \lambda D_x) + \alpha(R_y + \lambda D_y) + \beta R_w$$



Experimental Setup

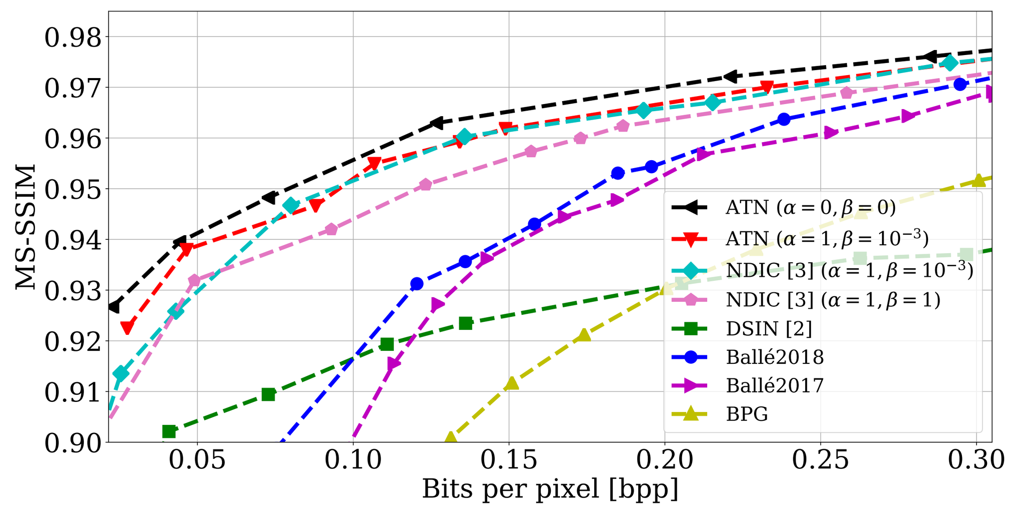
KITTI Stereo
(sync stereo)

Cityscape
(sync stereo)

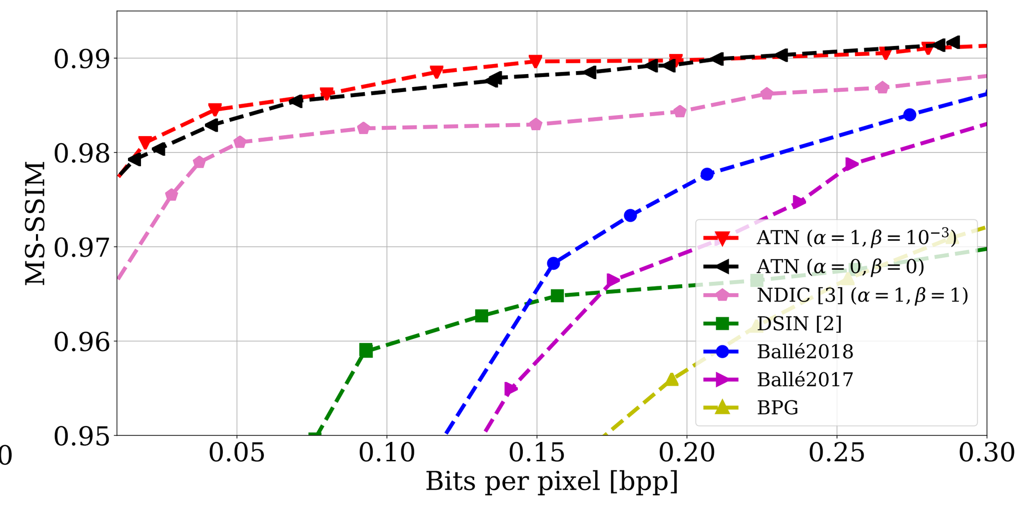
KITTI General
(unsync stereo)

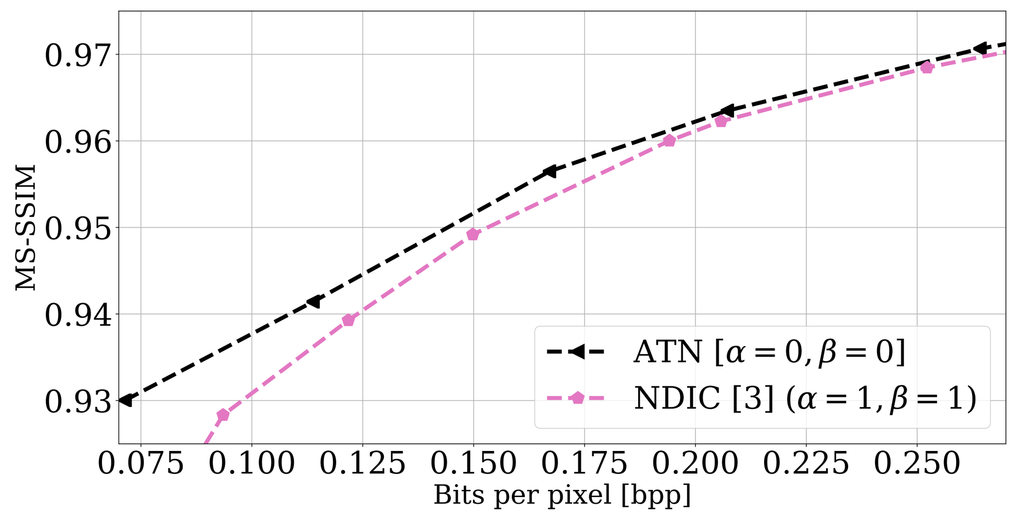
Rate-Distortion Performances

KITTI Stereo

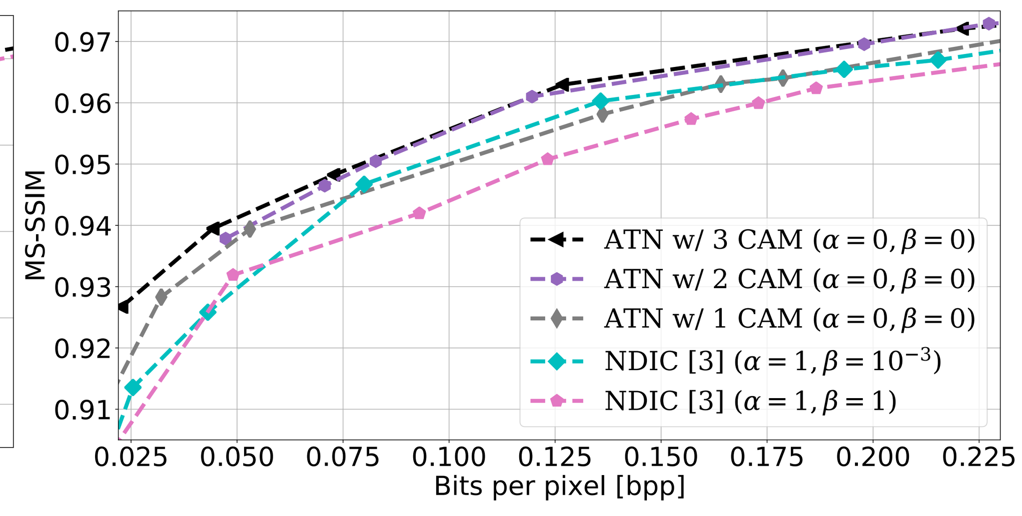


Cityscape





KITTI General

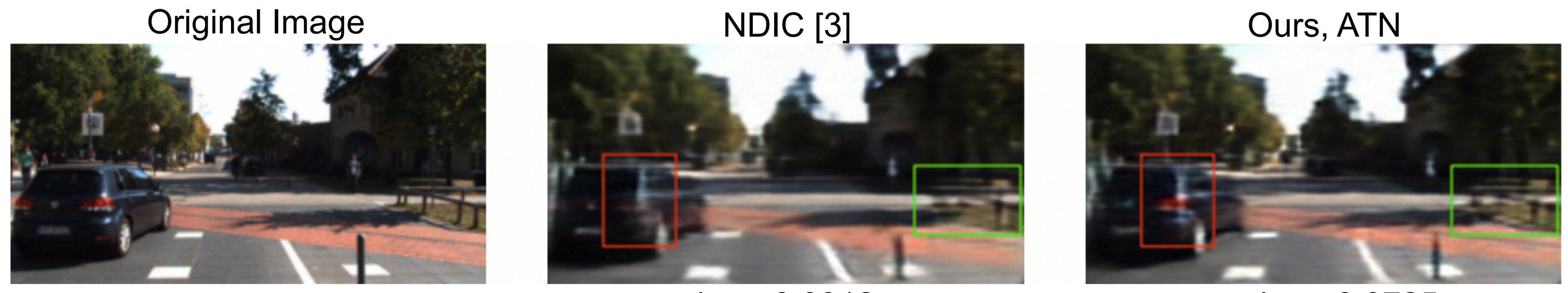


***Ablation** w/ KITTI Stereo*

Visual Examples

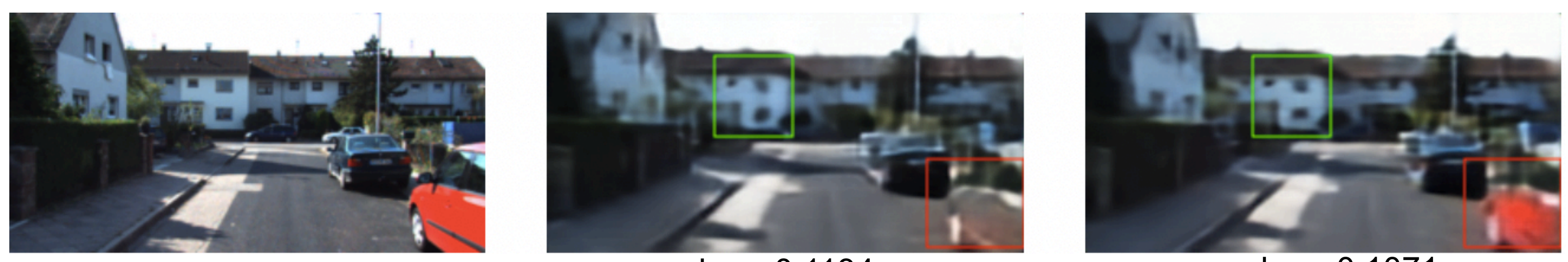
Synchronized stereo cameras

Original Image | NDIC [3] | Ours, ATN



bpp=0.0912 | bpp=0.0725

Unsynchronized stereo cameras

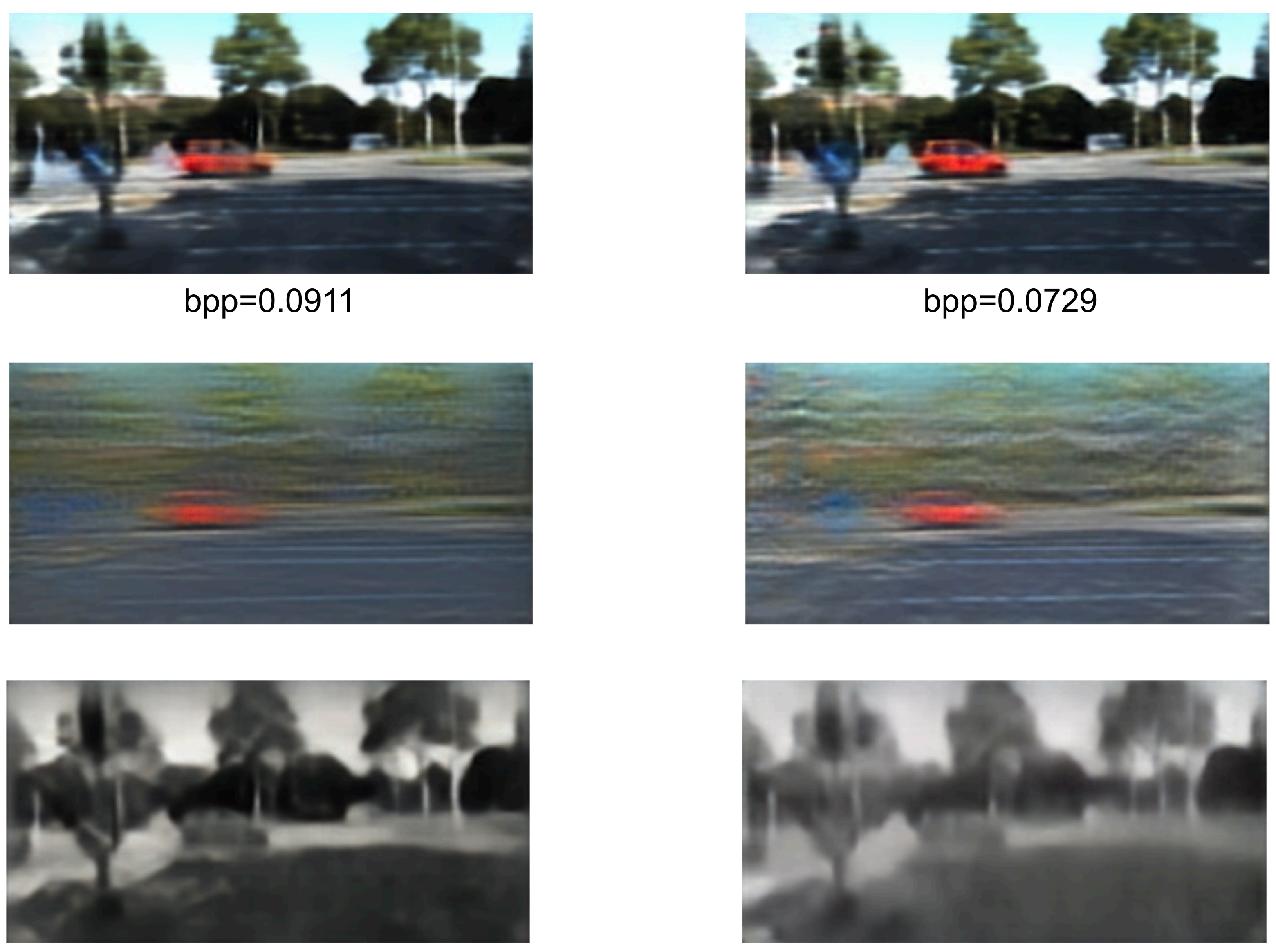


bpp=0.1134 | bpp=0.1071

Common and Local Information

NDIC [3] | Ours, ATN

Reconstructed Images | Common Information | Local Information



bpp=0.0911 | bpp=0.0729

References

[1] A. Wyner and J. Ziv, ``The rate-distortion function for source coding with side information at the decoder", *IEEE Trans. Inf. Theory*, 1976.

[2] S. Ayzik and S. Avidan, ``Deep image compression using decoder side information", *ECCV*, 2020.

[3] N. Mital, E. Ozyilkan, A. Garjani, and D. Gunduz, ``Neural distributed image compression using common information", *DCC*, 2022.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ``An image is worth 16x16 words: Transformers for image recognition at scale. In 9th International Conference on Learning Representations", *ICLR*, 2021.