Improving Feature Stability during Upsampling — Spectral Artifacts and the Importance of Spatial Context

Shashank Agnihotri[1], Julia Grabinski[1 2 3], Margret Keuper[1 4]
[1]Data and Web Science Group, University of Mannheim, Germany [2]Fraunhofer ITWM, Kaiserlautern
[3]Institute for Machine Learning and Analytics, Offenburg University of Applied Sciences [4]Max-Planck-Institute for Informatics, Saarland Informatics Campus, Germany

Motivation

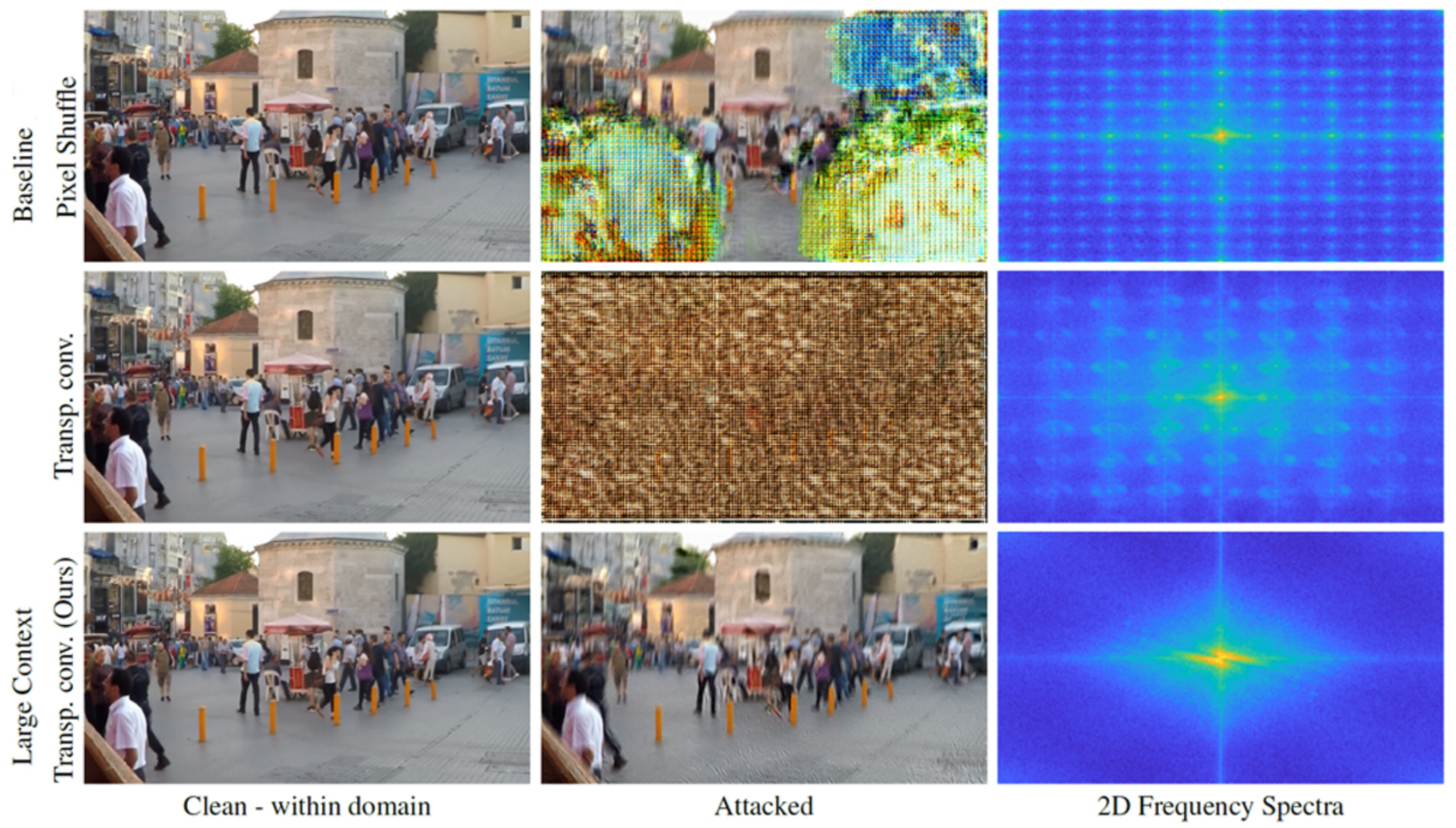

Figure 1. Image restoration using NAFNet [1] variants on GoPro images [2].

Introduction

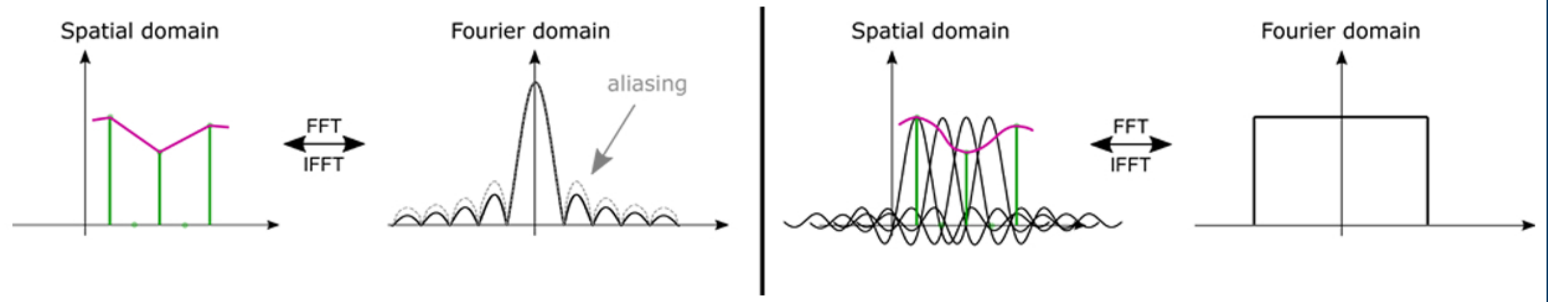

Figure 2. (Left) Linear interpolation (pink) of the sample (green) causes aliases. (Right) Optimal signal reconstruction (pink) is achieved by sinc interpolation. In practice our spatial context is limited and the interpolation function is discrete. Yet, increasing the kernel size enables the approximation of larger sinc-like structures.

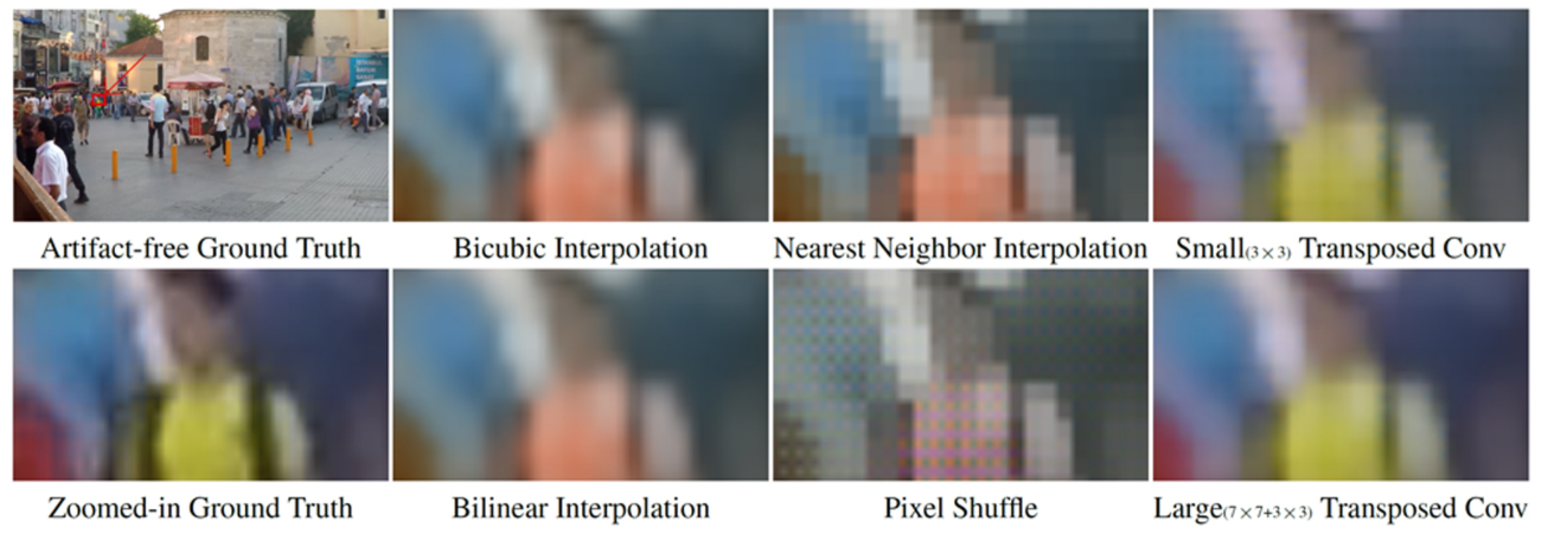

Figure 3. Images downsampled with 3x3 MaxPooling and then upsampled using various upsampling techniques.

Hypothesis

***Hypothesis 1: Large Context Transposed Convolutions (LCTC)* i.e. Large kernels in transposed convolution operations provide more context and reduce spectral artifacts and can therefore be leveraged by the network to facilitate better and more robust pixel-wise predictions.**

***Hypothesis 2 (Null Hypothesis H2):* To leverage prediction context and reduce spectral artifacts, it is crucial to increase the size of the transposed convolution kernels (upsample using large filters). Increasing the size of normal (i.e. non-upsampling) decoder convolutions does not have this effect.**

What causes spectral artifacts?

Consider, w.l.o.g. a 1D signal I and its DFT $\mathcal{F}(I)$ with k being the index of discrete frequencies

$$\mathcal{F}(I)_k = \sum_{j=0}^{N-1} e^{-2\pi i\cdot\frac{jk}{N}} \cdot I_j, \quad \text{for} \quad k = 0, \ldots, N-1.$$

During decoding, we upsample I to I^{up} with a factor of 2, for $\bar{k} = 0, \ldots, 2N-1$

$$\mathcal{F}(I)^{\mathrm{up}}_{\bar{k}} = \sum_{j=0}^{2N-1} e^{-2\pi i\cdot\frac{j\bar{k}}{2\cdot N}} \cdot I^{\mathrm{up}}_j = \sum_{j=0}^{N-1} e^{-2\pi i\cdot\frac{2\cdot j\bar{k}}{2\cdot N}} I_j + \sum_{j=0}^{N-1} e^{-2\pi i\cdot\frac{2\cdot(j+1)\bar{k}}{2\cdot N}} \bar{I}_j, \qquad (1)$$

the second term in (1) can be dropped in bed of nails sampling where $\bar{I}_j=0$
Since, the first term resembles $\mathcal{F}(I)$, we rewrite (1) using a Dirac impluse comb as,

$$(1) = \sum_{j=0}^{2N-1} e^{-2\pi i\cdot\frac{j\bar{k}}{2\cdot N}} \cdot \sum_{t=-\infty}^{\infty} I^{\mathrm{up}}_j \cdot \delta(j-2t). \qquad (2)$$

If we now apply pointwise multiplication with the Dirac impulse comb as convolution in the Fourier domain,

$$\mathcal{F}(I)^{\mathrm{up}}_{\bar{k}} = \frac{1}{2}\sum_{t=-\infty}^{\infty}\left(\sum_{j=-\infty}^{\infty} e^{-2\pi i\cdot\frac{j\bar{k}}{2N}} I^{\mathrm{up}}_j\right)\left(\bar{k}-\frac{t}{2}\right) \qquad (3)$$

$$\overset{(1)}{=} \frac{1}{2}\sum_{t=-\infty}^{\infty}\left(\sum_{j=-\infty}^{\infty} e^{-2\pi i\cdot\frac{j\bar{k}}{N}} \cdot I_j\right)\left(\bar{k}-\frac{t}{2}\right) = \frac{1}{2}\sum_{t=-\infty}^{\infty} \mathcal{F}(I)_{\bar{k}}\left(\bar{k}-\frac{t}{2}\right).$$

Such upsampling create high-frequency replica of the signal at $t/2$ for t in $-\infty, \ldots, \infty$ in $\mathcal{F}(I)^{up}$ and spatial frequencies apparent beyond array position $N/2$ will be impacted by spectral artifacts if no appropriate countermeasures are taken.

Methodology

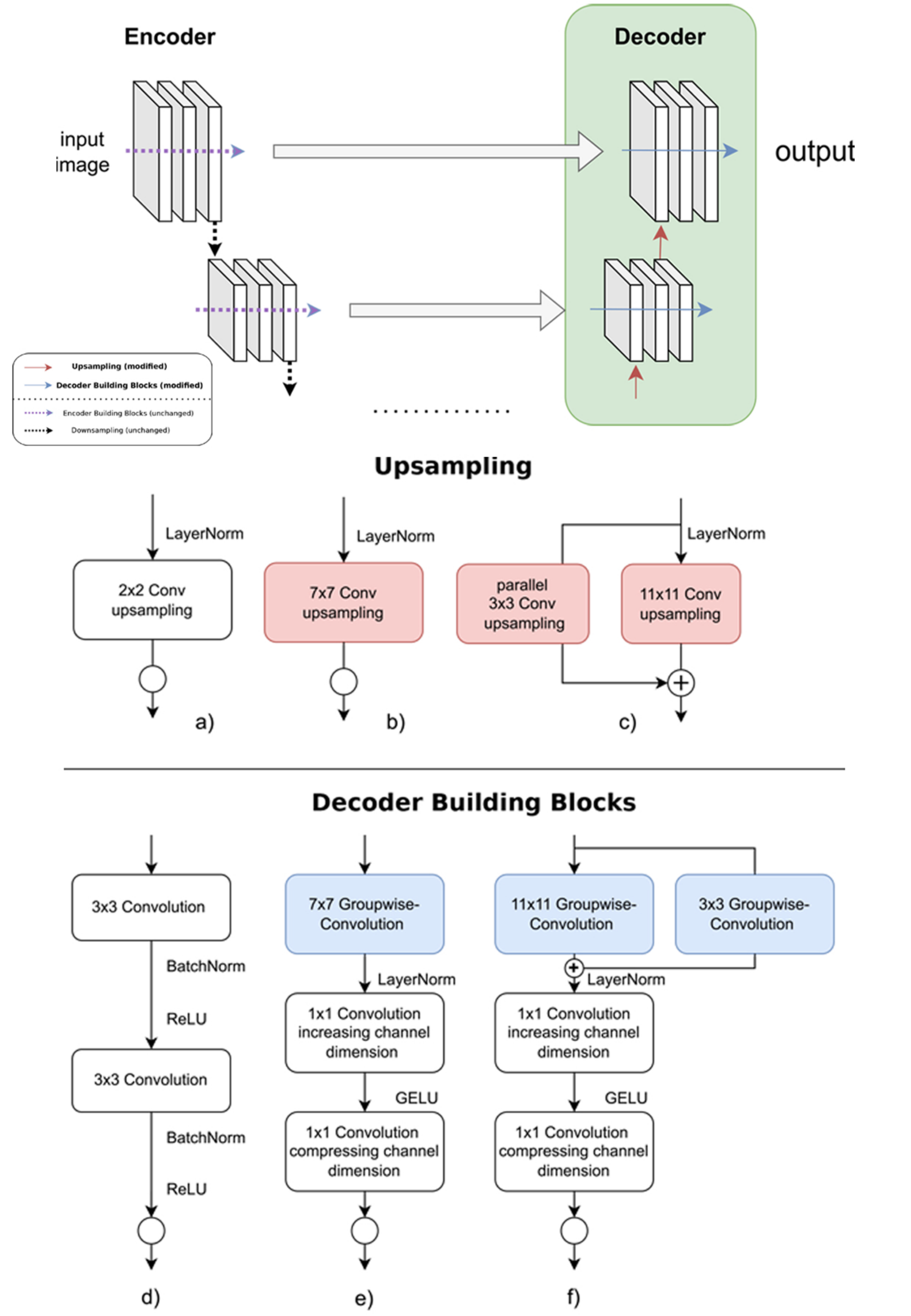

Figure 4. Abstract representation of the encoder-decoder architectural modifications. Our study focuses on the model decoder (in green).

Results

Figure 5. Restorations using NAFNet (uses PixelShuffle) and its variants including LCTC under PGD attack.

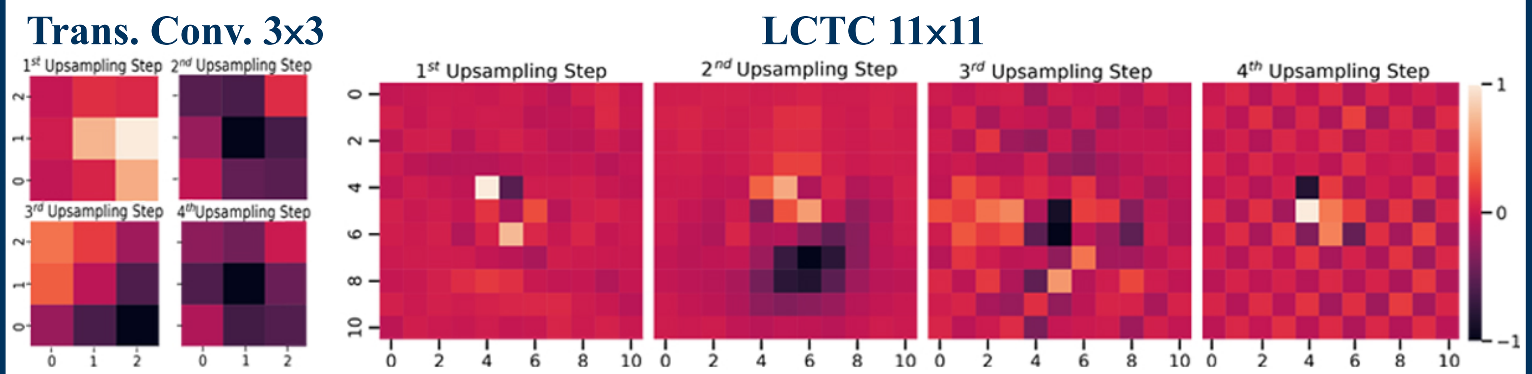

Figure 6. Normalized kernel weights from a random channel for the NAFNet models above.

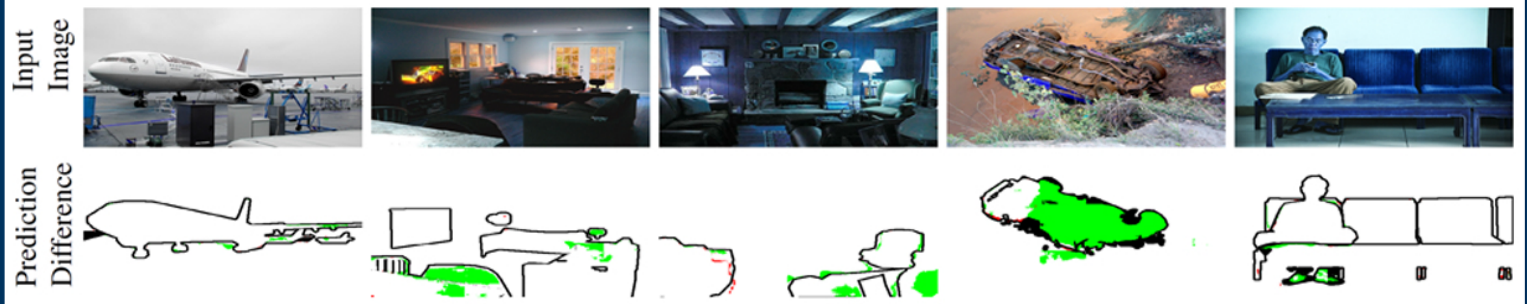

Figure 7. Difference in semantic segmenation mask predictions between LCTC and baseline transposed convolution using UNet [3] with ConvNeXt [4] encoder on PASCAL VOC2012 [5] images.

References and Useful Links

[1] Chen, L., Chu, X., Zhang, X., Sun, J. (2022). Simple Baselines for Image Restoration. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol 13667. Springer, Cham. https://doi.org/10.1007/978-3-031-20071-7_2
[2] Nah, S., Hyun Kim, T., & Mu Lee, K. (2017). Deep multi-scale convolutional neural network for dynamic scene deblurring. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3883-3891).
[3] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional networks for biomedical image segmentation." Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015.
[4] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11976-11986).
[5] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, (2012)

Paper

Margret Keuper acknowledges funding by the DFG Research Unit 5336 - Learning to Sense