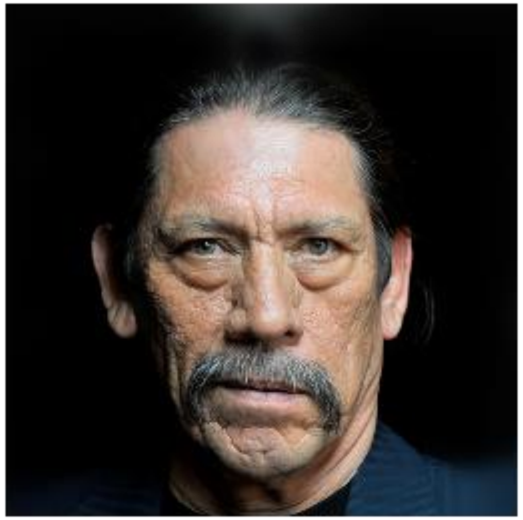
Generative Image Inpainting with Contextual Attention

MAXIMILIAN SCHLÖGEL AND PHILIPP LINTL

Machete without moustache?

Raw Image:



Input:

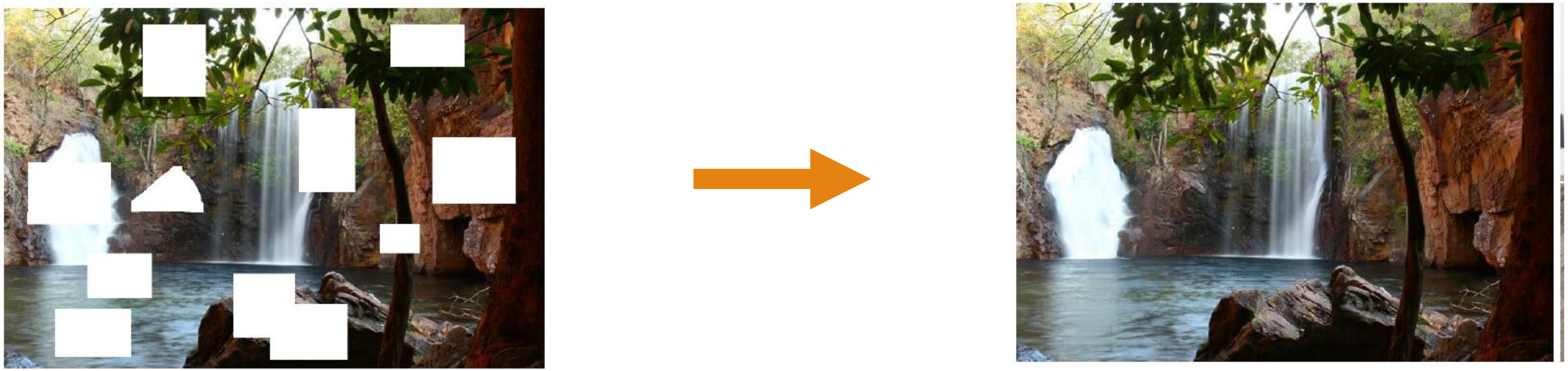
Our Result:

Structure

1. Introduction: Task and Applications
2. Related Work
3. Contribution of this Paper
   - Network Architecture
   - Contextual Attention
4. Experiments
5. Further Research
6. Outlook

Task: Image Inpainting

- Inpainting missing regions of image:



- Important for photo editing, image-based rendering, computational Photography
- Challenge: construction of realistic, semantically plausible pixel region
- Problems: distorted, blurry, inconsistent with surroundings and textures

Related Work

- **Diffusion:** propagates surrounding appearance into hole region
- **Patch based** : Filling hole by matching and stitching similar background patches

- Works for background filling on stationary images, small/thin missing regions
- **Problems:**

Slow, novel contents, high level semantics

⇨ Complex and non-repetitive structures (faces, objects) not properly inpaintable

Learning based

CNNs / GANs:

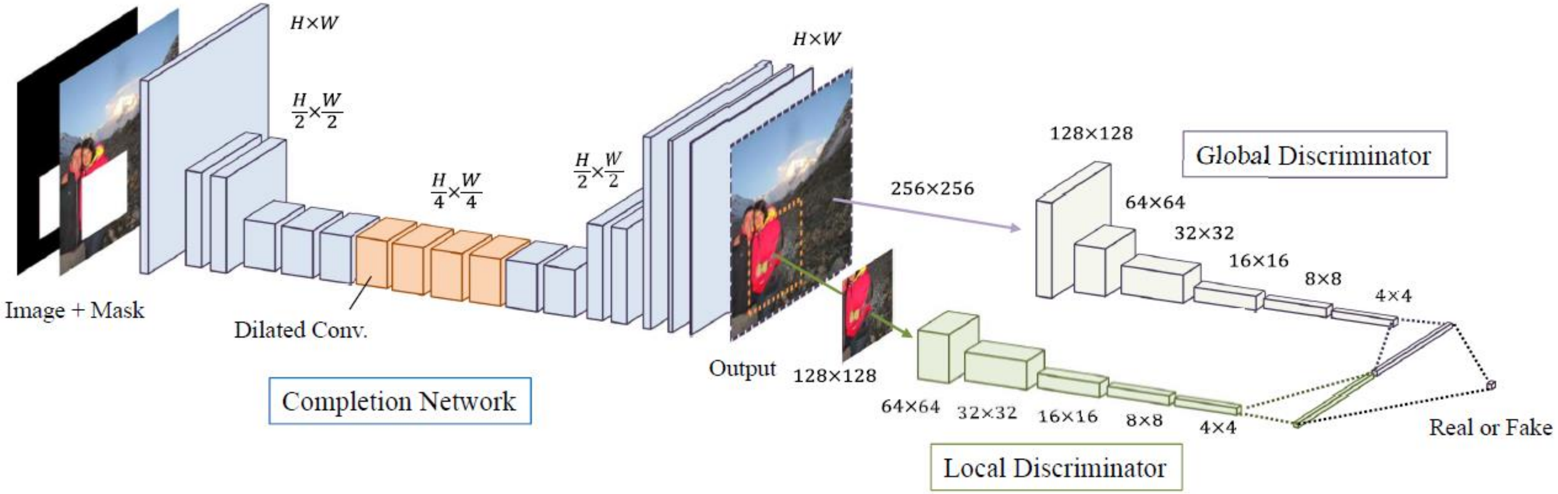
- Supervised Learning to predict missing region
- GAN: predict missing pixels by adversarial training (Discriminator: real image or completion)

Problems:

- Boundary artifacts (coherence between boundary and background not enforced)
- Distorted structures, blurry textures, novel structures

⇒ Convolutions: long term correlations between distant context and hole not covered

Main source: Iizuka et al (2017):



- Local Discriminator: Inpainted region locally consistent?
- Global Discriminator: Whole image consistent?
- Dilated Convolution increases receptive fields to allow more distant influence

Problems:

- slow (2 months of training), large holes not properly inpainted

Contribution

- Dilated convolutions right step, but symmetric grid -> no weighting of feature importance
- Improve information propagation from distant locations:

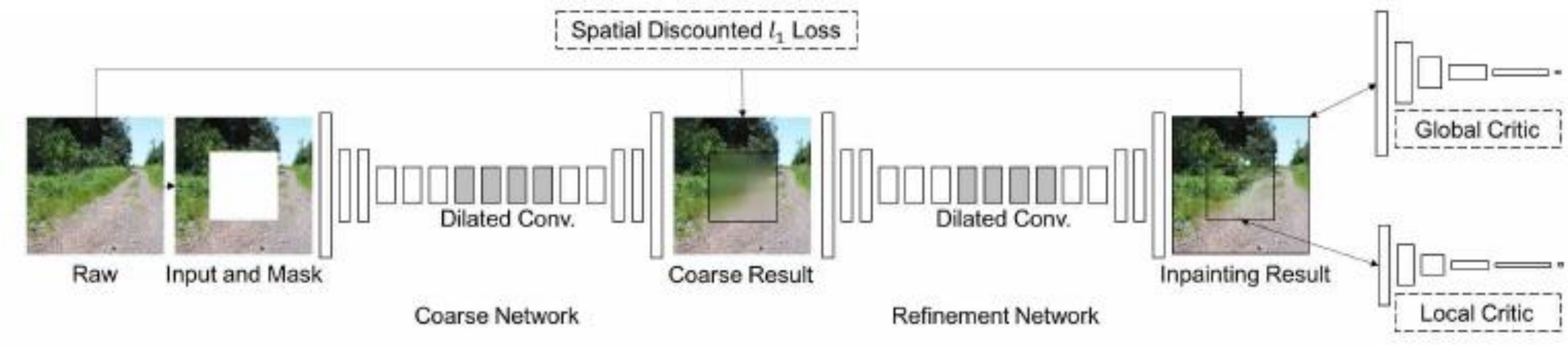
  To synthesize novel image structures and utilize surrounding image features

⇨ Two-stage Encoder-Decoder Architecture

⇨ Authors propose a novel **contextual attention layer**

⇨ Further adjustments to attend weaknesses of Iizuka et al.

Architecture



- 2 stage Coarse-to-fine architecture: 2 Encoder-Decoder
- Coarse prediction of hole region ⇨ input to refinement network
- Coarse: trained only on reconstruction loss
- Refinement: reconstruction **and** GAN losses

Architecture (2)

Wasserstein loss outperforms GAN loss for image generation task

$$L = \min_{G} \max_{D \in \mathcal{D}} \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_r}[D(\mathbf{x})] - \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathbb{P}_g}[D(\tilde{\mathbf{x}})] + \mathbb{E}[l_1(\mathbf{x}, \tilde{\mathbf{x}})]$$ + penalty term

- $\tilde{\mathbf{x}} = G(\mathbf{z})$, **z:** input mask to Generator, **x** sampled training patch
- One loss function per Discriminator
- Notation: pixelwise reconstruction loss $l_1(\mathbf{x}, \tilde{\mathbf{x}}) = \| G(\mathbf{z}) - \mathbf{x} \|$

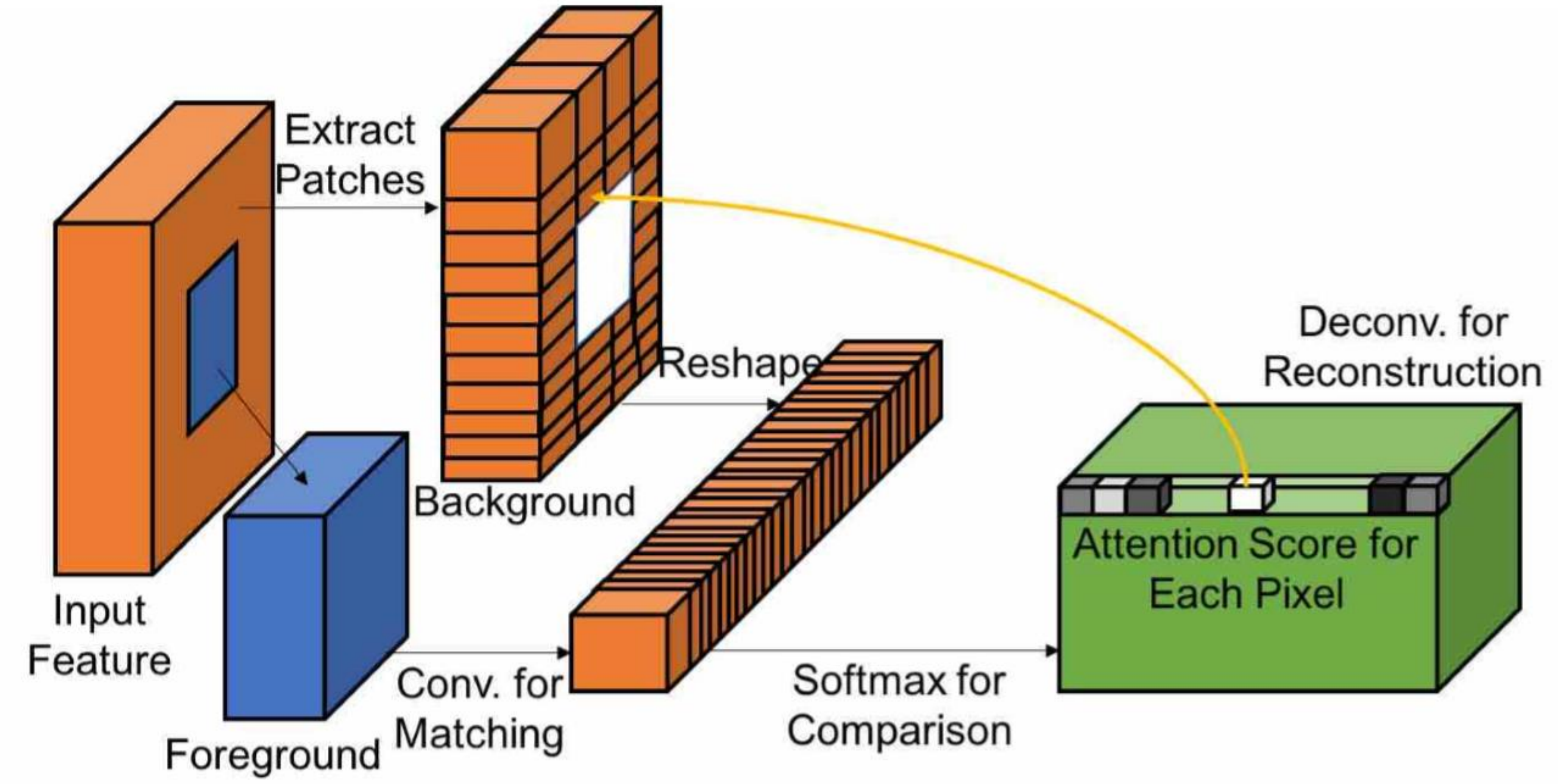
Spatially discounted reconstruction loss:

- Given a context and hole region: Many plausible solutions / possibly dissimilar to original image
- Strong reconstruction loss to ground truth: misleading towards training images
- Pixels closer to boundary have less ambiguity ⇒ higher weight (distant pixels less weight)

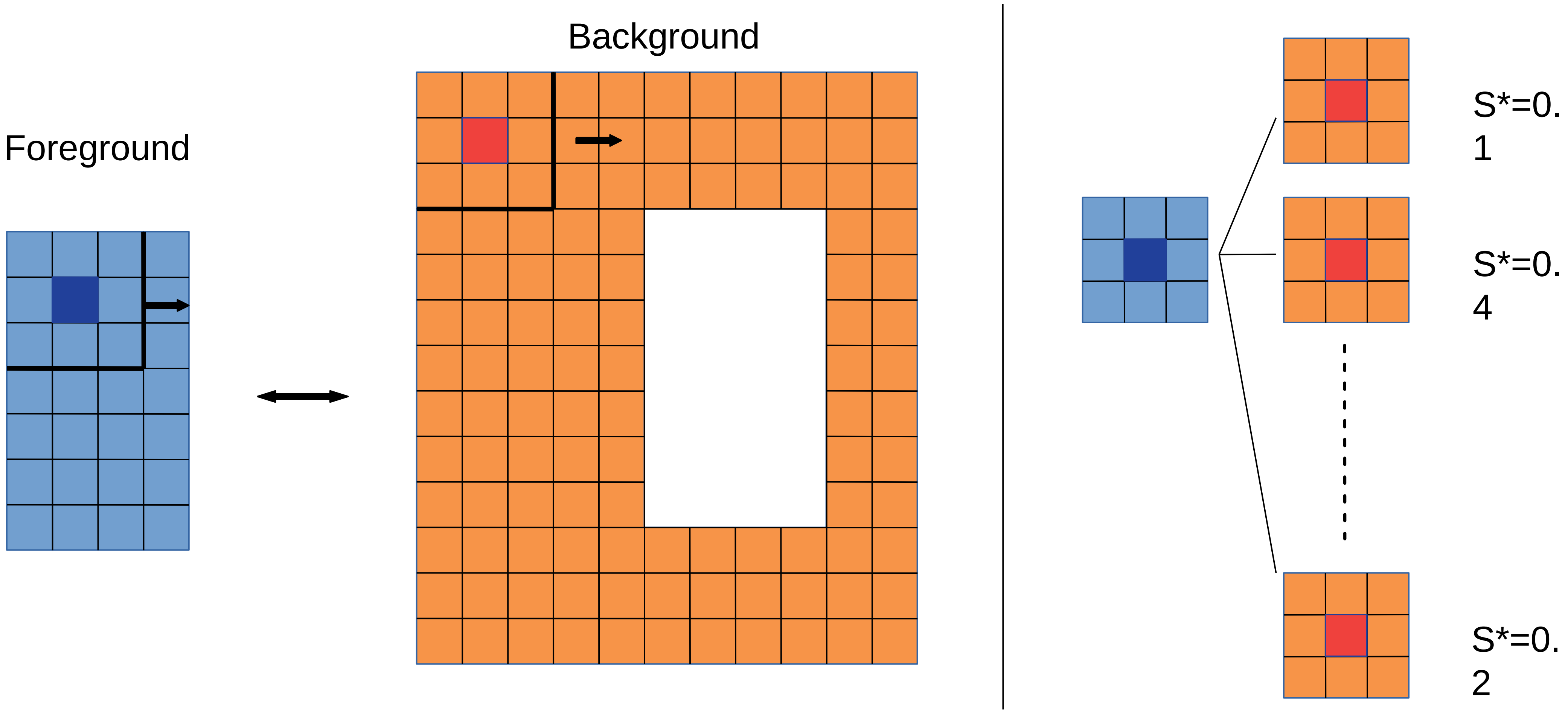
Contextual Attention Layer

- One Contribution of this paper is the **Contextual Attention Layer**
- It is applied for **generating finer image from the coarse prediction** (see above).
- Solves short-sightedness of ConvNets, by learning how to find similar or relevant image features

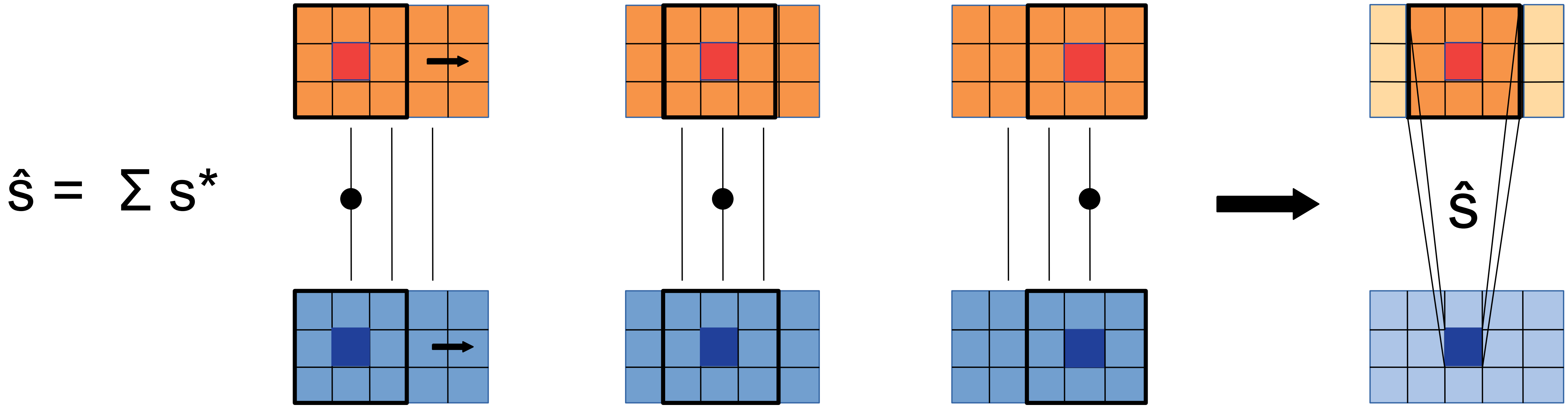
Contextual Attention Layer



Contextual Attention Layer



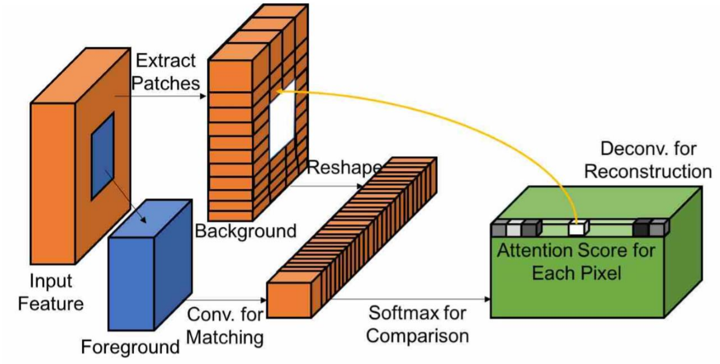
Attention Propagation



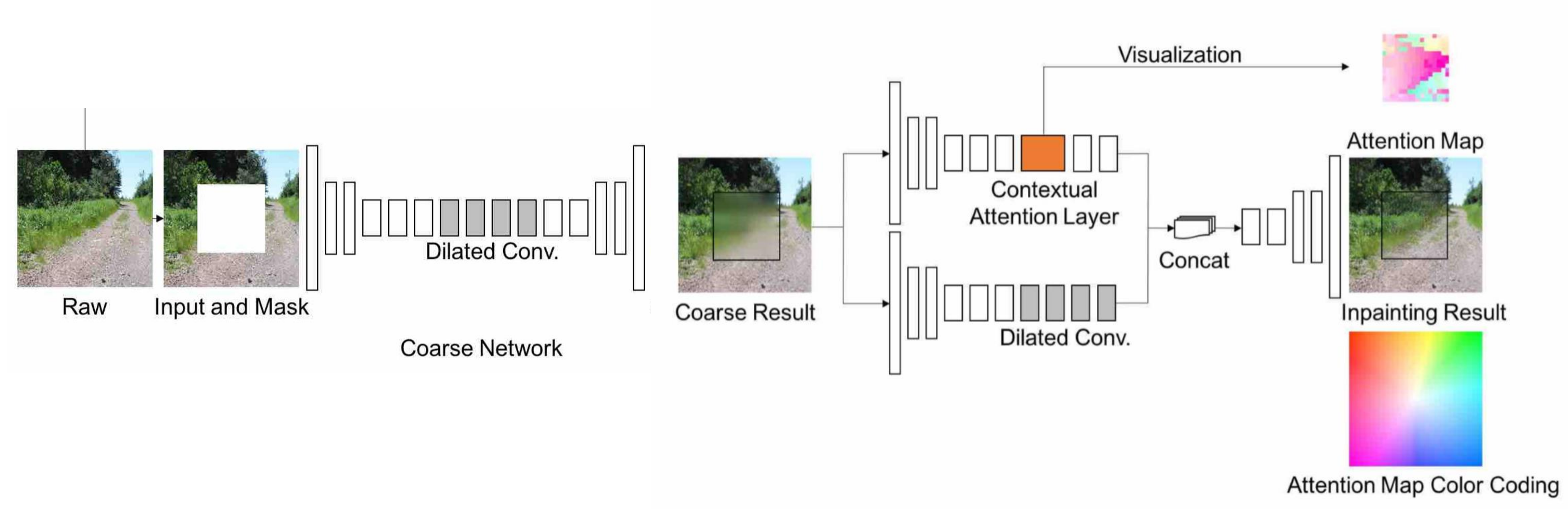
Contextual Attention Layer

- The contextual attention layer learns where to borrow feature information from, by:

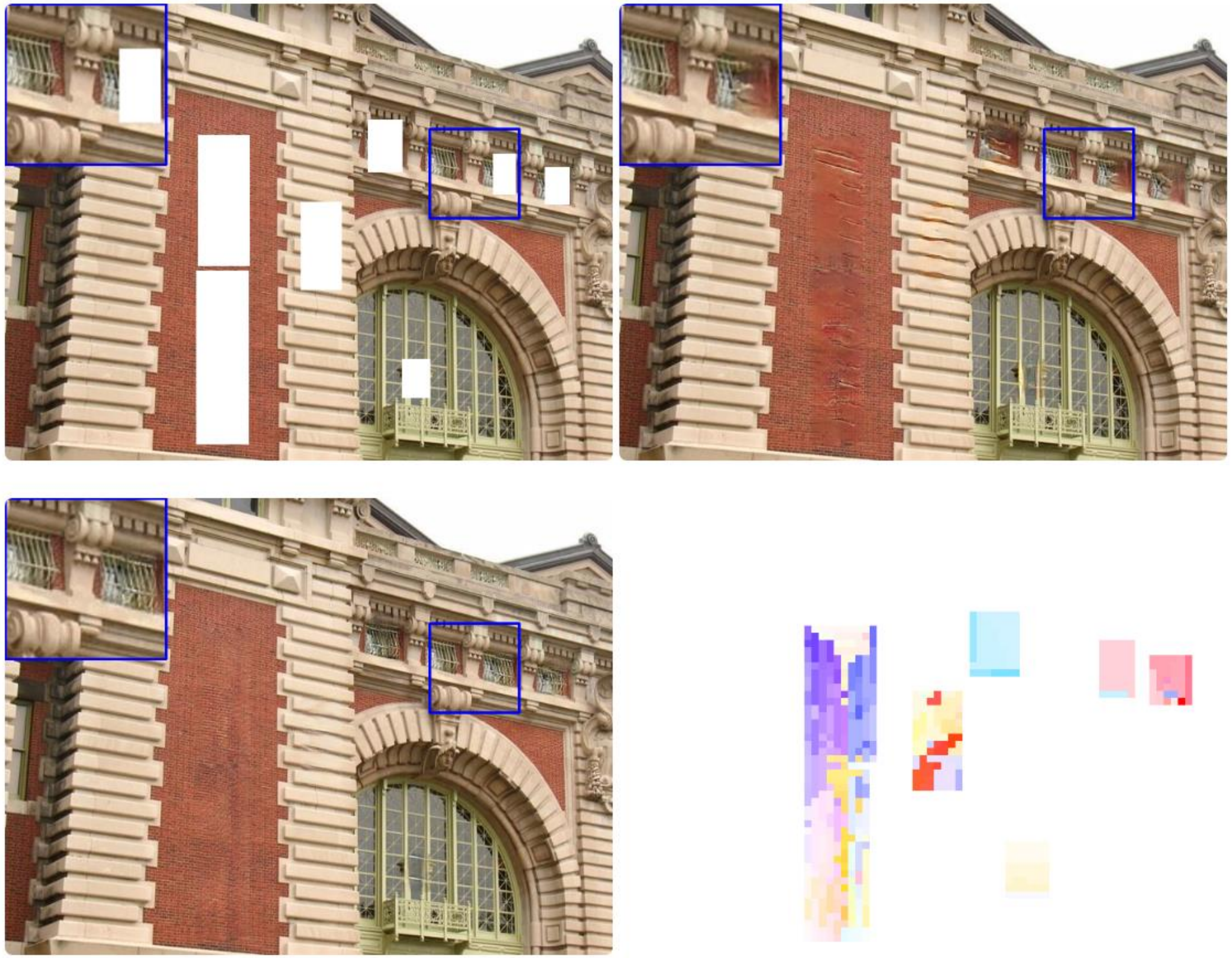
1) Translating each background patch into a **convolutional filter** (channel-wise)
2) Compute **cosine-similarity** between those background patches and patches around pixel in the coarse foreground of the firt NN-part
3) Use **attention propagation**
4) Apply softmax → get **best match** and use it for deconvolution to generate the inpainting resullt.



Unified Inpainting Network



Experiments – Qualitative Comparison



Experiments – Qualitative Comparison

Original with Mask

Baseline

Our Model

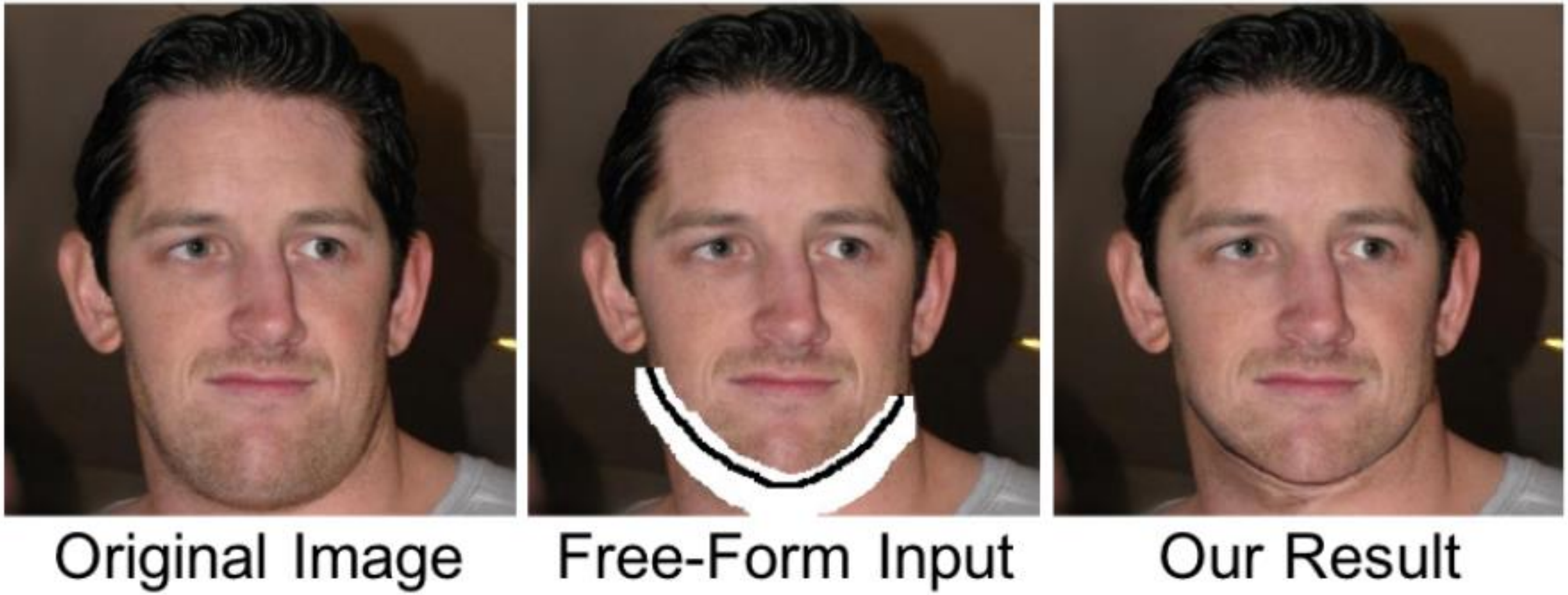
Experiments – Quantitative Comparison

- Image Inpainting Tasks lack good (quantitative) metrics
- Paper uses L-1, L-2, and
- PSNR (Peak-Signal-to-Noise ration)
- TV (Total Variation)

Method	ℓ_1 loss	ℓ_2 loss	PSNR	TV loss
PatchMatch [3]	16.1%	3.9%	16.62	**25.0%**
Baseline model	9.4%	2.4%	18.15	25.7%
Our method	**8.6%**	**2.1%**	**18.91**	25.3%

Outlook

- Cited by 62 paper
- Follow-up Paper from same authors: Free -Form Image Inpainting with Gated Convolution



Interesting Findings

Input:

Raw Image:

Our Result:

Interesting Findings

Input:

Raw Image:

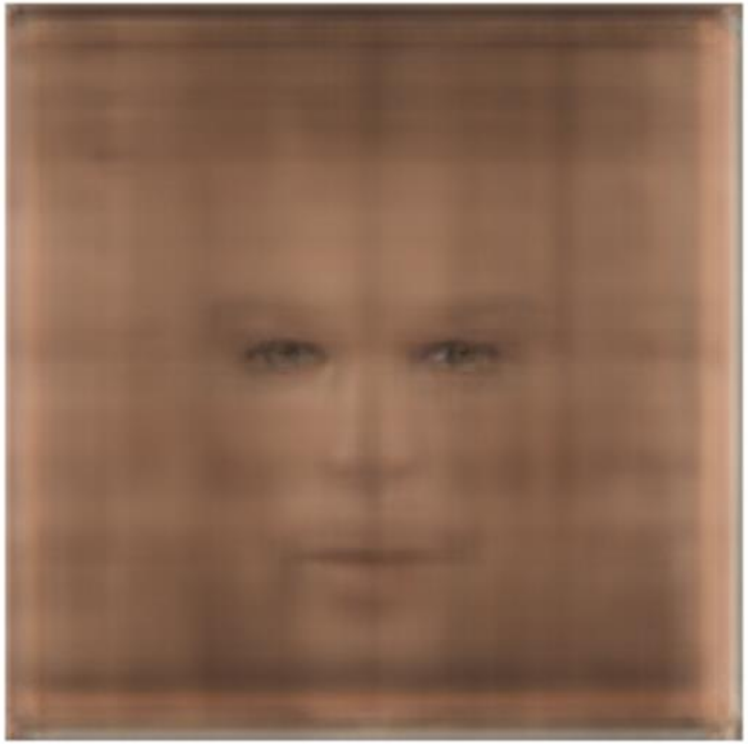
Our Result:

Interesting Findings

Input:

Raw Image:

Our Result:



Weaknesses

- Memory cost of contextual layer
- No benchmark dataset so far
- No agreed general performance measures

Strengths

- No postprocessing necessary to yield proper inpaintings
- Explainable (visualization of generation process)
- Self-supervised setup
- Training time (1 week vs. 2 months)

Summary

- Enables distant feature borrowing
- Contextual Attention improves inpainting quality especially for larger holes
- Architecture adjustments increase training stability and speed (1 week vs. 2 months)

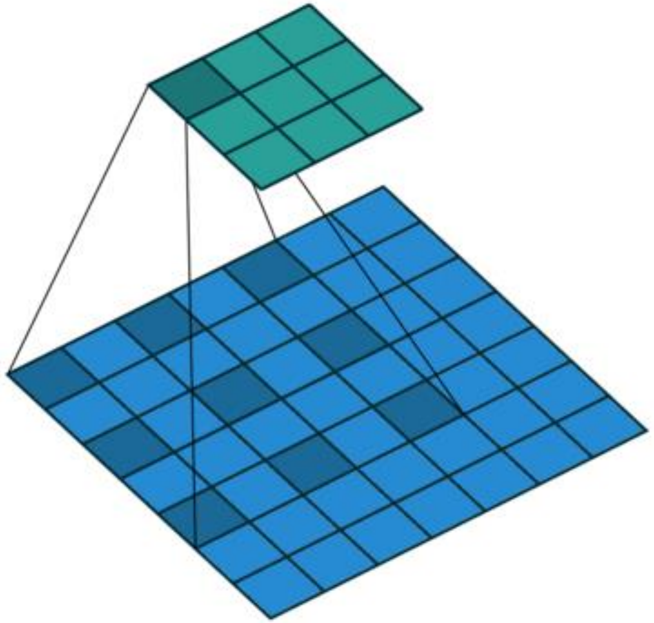
Criticism of Paper

Lack of details:

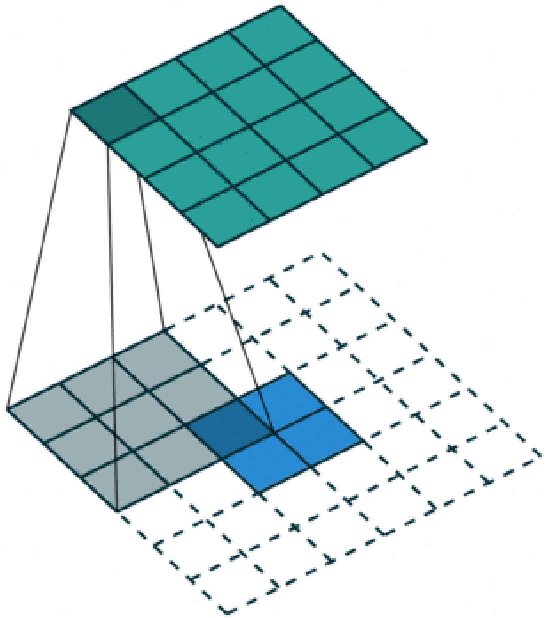
- Full loss function: 2 adversarial losses, reconstruction loss but no complete Loss function
- Training procedure: End-to-end and pseudo-algorithm not very understandable
- Generator / Discriminator details more comprehensible in related work
- Iizuka 2017 provide much more detailed descriptions and formulas

Appendix

Dilated Convolution



Deconvolution



Source of the gifs: https://github.com/vdumoulin/conv_arithmetic