# Video DeCaptioning using U-Net with stacked dilated convolutional layers

Shivansh Mundra, Arnav Kumar Jain*, Sayan Sinha*

**Abstract** We present a supervised video decaptioning algorithm driven by an encoder-decoder pixel prediction. By analogy with auto-encoders, we propose U-Net with stacked dilated Convolution layer  a convolutional neural network trained to generate the decaptioned version of an arbitrary video with subtitles of any size, colour or background. Also, our method doesn't require mask of the region with text to be removed. In order to succeed at this task, our model needs to both understand the content of the entire frames of video, as well as produce a visually appealing hypothesis for the missing part behind text overlay. When training with our model, we have experimented with both a standard pixel-wise reconstruction loss, as well as total variation loss. The latter produces much sharper results because it enforces local inherent nature in generated image. We found that our model learns a representation that captures not just appearance but also the semantics of visual structures. We quantitatively demonstrate the effectiveness of including dilated convolution layers and residual connections in bottleneck layer in reconstruction of videos without captions. Furthermore, our model be used for semantic inpainting tasks, either stand-alone or as initialization for non-parametric methods.

Shivansh Mundra
Department of Mechanical Engineering, Indian Institute of Technology Kharagpur, e-mail: shivanshmundra1@gmail.com

Arnav Kumar Jain
Department of Mathematics, Indian Institute of Technology Kharagpur e-mail: arnavkj95@gmail.com

Sayan Sinha
Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur e-mail: sayan.sinha@iitkgp.ac.in

* Denotes equal contribution.

## 1 Introduction

Video decaptioning refers to the task of replacing the text overlays in frames with semantic coherent regions. In this work, we explore the application of state-of-the-art computer vision algorithms to address the challenge in an automated fashion. The task requires first finding the region with captions and then predicting the high level context, hence making it significantly more difficult when compared to classical image or video inpainting methods. However, decaptioning becomes increasingly more difficult, when the subtitles cover most of the parts of the frame and are of different size, font and colours.

Videos often have captions embedded into them such that one is unable to turn them off when not required. Therefore, even if it is more comfortable viewing the video without the captions, there is hardly any way out. We extend the fully convolutional network [1] and modify its architecture such that when it is fed with videos having captions, it tries to predict what they would look like had the captions not been there. The main idea lies in the fact that we propose an encoder-decoder model which supplements the usual contracting network by successive layers, where pooling operators are replaced by up-sampling operators. Hence, these layers help in increasing the resolution.

## 2 Related Work

Initially, major works of inpainting could be categorised into three groups of works. In the works of Hirani and Totsuka [2], frequency and spatial domain information is blended to fill in a given region with a selected texture. Disocclusion was another popular method introduced by M. Nitzberg et al. [3]. Thus it can be seen that non-learning approaches to image inpainting rely on propagating appearance information from neighboring pixels to the target region. They are specific to image sets and can be used to fill in only small regions of gaps.

Computer vision has made tremendous progress on semantic image understanding tasks such as classification, object detection, and segmentation in the past decade. Conventional Sparse coding methods [4] were sensitive to image orientation and environment and couldn't be generalized into cross domain works. Recently, Convolutional Neural Networks (CNNs) [5], have greatly advanced the performance in these tasks. The success of such models on image classification paved the way to tackle harder problems, including unsupervised understanding and generation of natural images.

Deep learning based methods typically initialise the gaps with values such as a constant or mean pixel value after which the resultant is passed through a deep convolutional network. In our paper, such an effort is not required as the captions are atop the image and are to be fed directly. D Pathak et. al[6] first introduced the concept of image inpainting using an encoder decoder netowork , adapting concepts

from autoencoders. In this paper we extend this concept to build an innovative model for the same.

We briefly review the related work in each of the sub-fields pertaining to this paper.

### 2.1 GAN based inpainting methods

Some recent work [7] [8] have shown convincing results in patch based inpainting. They used Generative Adversarial Networks (GANs) [9] in two contexts, one global discriminator and the other one is local discriminator. But GAN based methods often fail when it comes to inpainting on data set with diverse classes. Hence these methods couldn't be directly applied in the task of decaptioning. However, [10] showed that GAN are able to produce more visually sharper and pleasing images and leading to the use of loss of Discriminator in addition to the normal losses to produce sharper results.

### 2.2 AutoEncoder based inpainting methods

Works like [11] has shown that Auto Encoder-Decoder based methods have produced good results in image denoising and image inpainting tasks. Also, this work has shown that the shape of the mask (region that needs to inpainted) is not required to be given as input to the model. They directly take the corrupted frame and output the reconstructed images. We also use an auto-encoder in our case, since the region with subtitles can occupy different areas in frames.

### 2.3 Unsupervised methods for Inpainting

Feature Learning methods such as [6] have shown good results on high resolution images where a large section of an image was needed to be inpainted. They have trained on reconstruction and adversarial losses which resulted in real looking images and closer to manually inpainted image. One good advantage of learning features of the image is to understand it's semantics which is important for unsupervised inpainting.

## 3 Methodology

We propose an end-to-end training method for the purpose of video decaptioning. Our purpose of video decaptioning has been broken down such that we try to focus on regeneration of the entire image from the input, sans the captions. An encoder-decoder network suits our case the most. We focus more on learning representations than on feature based learning. The input is mapped to a set of feature representations which remain hidden as a part of the neural network black box model. These features are processed further to obtain image representation which have successfully overwritten the captions. We have an Auto-Encoder model for this task.

Residual networks [12] have been shown to be easy to train. We make use of this

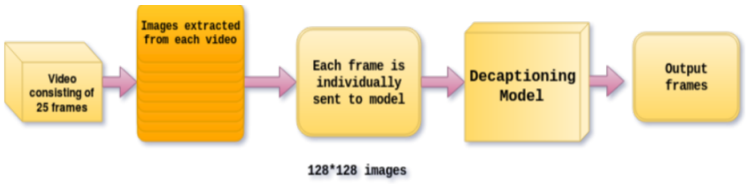


**Fig. 1** Training Workflow

concept to build a more robust model. Also, since we are making use of the model for a translation, there is a need to preserve features from earlier layers. The output mostly preserves the original image frame just replacing the captions, hence preserving certain features of the earlier layers is intuitively helpful.

We chose U-Net as the basic model on top of which we improvised. We tried incorporating certain elements from other deep learning frameworks into our model. These include the following:

- **Dilation:** The idea is to improve the receptive fields of the convolutions so that more learning can be involved in less number of layers. This is easily achieved using pooling which is a part of the original U-Net model. But, dilated convolutions help have a greater receptive field without compromising on resolution. Hence, incorporation of dilation was deemed important. [13]

- **Residual Skip Connections:** Deep networks are often difficult to train. In fact, a deeper network might not perform better than its shallower counterpart. Gradients get stalled and the error is larger. In order to make it easy to train such networks and to get over the issue of vanishing gradient, residual skip connections are often a good choice [14]. In order to enable it to learn the deviations, this acted as a great booster to our model

# 4 Model

We now introduce our solution architecture model U-Net with variants: CNNs that replace text overlay in frame of video to semantically coherent patch such that the scene is consistent spatially. We first give an overview of the general architecture, then provide details on the learning procedure and finally present various strategies for text region removal.

## 4.1 U-Net: Encoder-Decoder Pipeline

The overall architecture is a simple encoder decoder structure. Encoder takes captioned image frame as input and produces a latent representation while preserving semantic representation at each down-sampling layer which will be later used in decoder part. Decoder takes latent representation after it get convoluted from residual layers and up-sample the latent representation to reconstruct the frame without captions/text overlay. Decoder also incorporates different feature representations from encoder part while down-sampling because it content semantic information about original frame.

**Encoder** Our encoder is derived from pix2pix [15] which is also state of the art model in the context of image-to-image translation. We used five convolutional layers, each reducing width and height by a factor of 2 and increasing channel length by same factor. We used stride factor as 2 for above transformation. After each above operation, we used convolutional layer with stride 1 to connect feature maps together. These convolutional layers resulted to "bottleneck" layer with dimensions 4*4*512 = 8,192.

As this is not autoencoder, we need sufficiently large parametric representation of encoded image. However, if the encoder architecture is limited only to regular convolutional layers, there is no way for information to directly propagate from one corner of the feature map to another. This is so because convolutional layers connect all the feature maps together, but never directly connect all locations within a specific feature map.

So, we introduce stack of dilated Convolutional Layer in encoder part with dilation factor as (2,4,8,16). Dilated convolutional layers helps in generating more visually appealing outputs.

**Decoder** We now discuss the second half of our pipeline, the decoder, which generates pixels of the image using the encoder features. The channel-wise fully-connected layer is followed by a series of five up-convolutional layers [16] [17] [18] with learned filters, each with a rectified linear unit (ReLU) activation function. A up-convolutional is simply a convolution that results in a higher resolution image. It can be understood as upsampling followed by convolution (as described in [16]), or convolution with fractional stride (as described in [17]). We have introduced skip connections between symmetrical layers of encoder decoder so as to add feature semantic information in decoder in the reconstruction process. The intuition behind

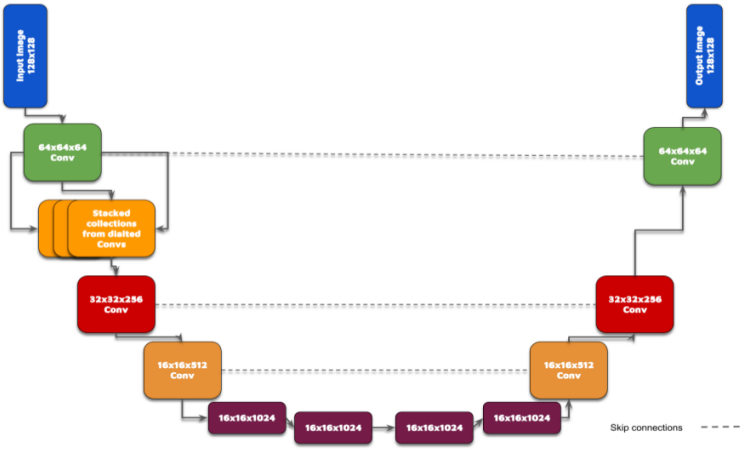


**Fig. 2** Model Architecture

this is straightforward: Unlike auto encoder, here we need to remove text overlay from a scene. Encoder just compresses the frame into a low-dimensional encoding while ignoring finer textures in its features. So we need to incorporate encoded feature representation. This is like feeding VGG feature representation to decoder. The series of up-convolutions and non-linearities comprises a non-linear weighted upsampling of the feature produced by the encoder until we roughly reach the original target size. The architecture of the U-Net adopted in this work is shown. However, U-Net based architecture is consistent with the poor performance of CNN-based inpainting in recovering fine textures. [19]

## 4.2 Dilated Convolutions

Unfortunately, these CNN-based methods often create boundary artifacts, distorted structures and blurry textures inconsistent with surrounding areas. We found that this is likely due to ineffectiveness of convolutional neural networks in modeling long-term correlations between distant contextual information and the hole regions. For example, to allow a pixel being influenced by the content of 64 pixels away, it requires at least 6 layers of 3 Œ 3 convolutions with dilation factor 2 or equivalent [17, 42]. So to recover complex image semantics and structures, we introduce a stack collections of dilated convolutions layer in the encoder part. Our architecture uses dilated convolutions to systematically aggregate multi-scale contextual information without losing resolution(which is already less in the given dataset). The architecture is based on the fact that dilated convolutions support exponential expansion of the receptive field without loss of resolution or coverage. [20] The encoder part is a

simple CNN with dilated convolutional layer trained with decoder part optimized on reconstruction loss to inpaint out the missing contents behind text overlays.

### 4.3 Residual Connections in bottleneck layer

After we incorporated U-Net and dilated Convolutions in our model architecture, we got convincing results but resolution was poor compared to ground truth. This was intuitively expected as simple Encoder-Decoder architecture trained on MSE loss generalize to give blurred output. The transposed convolutional (deconvolution) layer can learn up-sampling kernels, however, the process is similar to the usual convolutional layer and the reconstruction ability is limited. To obtain a better reconstruction performance, the transposed convolutional layers need to be stacked deeply, which means the process needs heavy computation. So, we propose skip connections as in bottleneck layer of encoder decoder architecture. As with typical Residual learning networks, the model is made to focus on learning residual output and this greatly helps learning performances in direction of increasing resolution, even in cases of shallow models. [21]

## 5 Results

As you can see in above images from test case, there is very little difference visible between images generated by adding dilated convolution layers because resolution of the data set provided in the challenge but we can see difference in the losses computed. We have kept residual connections in both the part of our experimentation. In our testing pipeline, we used a pre-trained model provided by organizers as a part of the baseline. First image was divided into 16 equal parts and each part was feeded to pretrained model to check if their was text overlay in the corresponding image. If there was text overlay in the part, it was replaced by similar part from predicted image from model. If score of text classification was below a threshold score, it was replaced by the corresponding part from the input image. This process was similar to Poisson Blending [22]. Our method took lesser time too reach optimal minimal compared to GAN based methods as there were no generator and discriminator trying to optimize simultaneously by min-max strategy. Also, our solution doesn't require binary mask for inpainting hence decreasing inference time. Our method took approximately 5 seconds to generate a decaptioned video.

As this was the first attempt in the field of video decaptioning, there weren't many baselines we could refer to. Hence we had shown comparison with Baseline and our model without dilated convolution. With this approach we came **2nd** in training

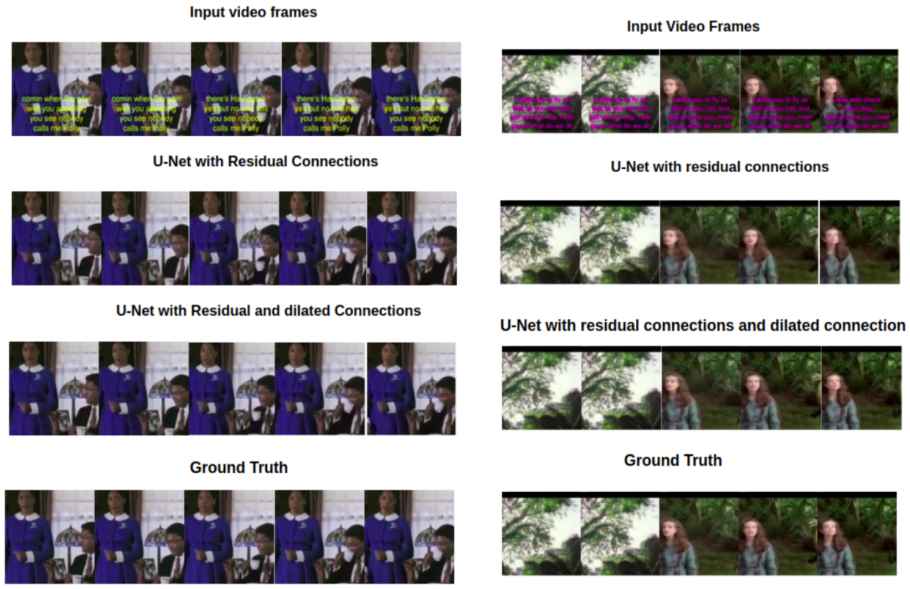


**Fig. 3** Left: Test Case 1, Right: Test Case 2

phase and **4th** in test phase of Chalearn Video Decaptioning Challenge.

**Table 1** Results

| **Method** | **MSE Loss** | **PSNR Loss** | **DSSIM Loss** |
|---|---|---|---|
| Baseline | 0.0022 | 30.1856 | 0.0613 |
| U-Net without dilation | 0.014 | 32.850 | 0.0511 |
| **U-Net with dilations** | **0.0012** | **32.1713** | **0.0482** |

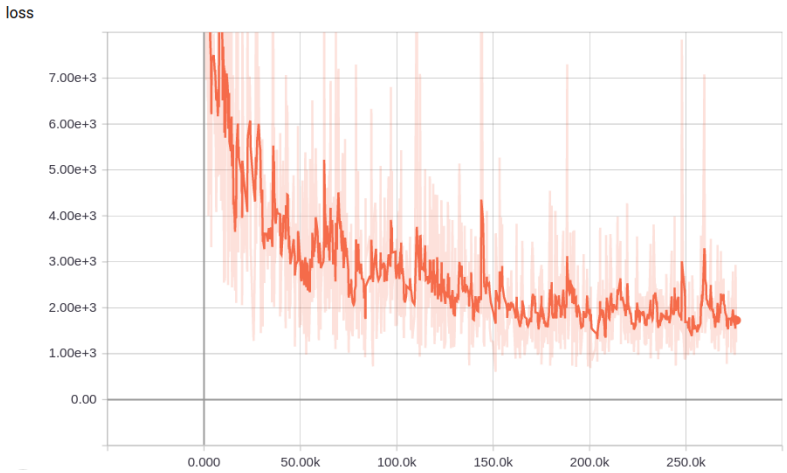


**Fig. 4** MSE Loss vs Iterations

## 6 Conclusion

From our experience in this competition, we came to following conclusion in the task of Video Decaptioning and related problem statements:

- Simple Auto Encoder-Decoder based solution are not good when it comes to noise removal from a large section as model is generating image from just encoded latent representation.
- Hence we need a model which have incorporated image semantics in the part of encoding and can be used while generating decaptioned image. U-Net based model was proven a good choice in the related field as it included skip connections between symmetric layers in the encoder-decoder part.
- As we needed to capture end to end semantics in the image to get global feature, we used stacked dilated convolution layer to incorporate global semantics in the encoding part. Here noise removal was to be done considering generated image was supposed to look real and dilated convolution layers were useful to that.
- resolutionencodericoder generally decrease the sharpness and resolution in the image generated, residual connections were added to improve sharpness. Although advantage of adding residual connection was not adding significant difference but it could increase resolution and visual appearance by significant margin when it comes to high resolution data set.
- We did not extract explicit mask for the region of text removal as it is implicitly learned by the encoder-decoder model.
- We didn't explore effect of temporal dimension in the process of video denoising but incorporating temporal dimension should help.

## References


1. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
2. Anil N. Hirani and Takashi Totsuka. Combining frequency and spatial domain information for fast interactive image noise removal. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 269–276, New York, NY, USA, 1996. ACM.
3. Mumford David Shiota Takahiro Nitzberg, Mark. Filtering, segmentation and depth. In *Springer-Verlag*.
4. Muhammad Hanif, Anna Tonazzini, Pasquale Savino, and Emanuele Salerno. Sparse representation based inpainting for the restoration of document images affected by bleed-through. In *Multidisciplinary Digital Publishing Institute Proceedings*, volume 2, page 93, 2018.
5. Kunihiko Fukushima. Neural network model for a mechanism of pattern recognition unaffected by shift in position-neocognitron. *IEICE Technical Report, A*, 62(10):658–665, 1979.
6. Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.
7. Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018.
8. Xinshan Zhu, Yongjun Qian, Xianfeng Zhao, Biao Sun, and Ya Sun. A deep learning approach to patch-based image inpainting forensics. *Signal Processing: Image Communication*, 2018.
9. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
10. Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2, page 4, 2017.
11. Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *Advances in neural information processing systems*, pages 341–349, 2012.
12. Fisher Yu, Vladlen Koltun, and Thomas A Funkhouser. Dilated residual networks. In *CVPR*, volume 2, page 3, 2017.
13. Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107, 2017.
14. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
15. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.
16. Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1538–1546, 2015.
17. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
18. Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.
19. Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017.
20. Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

21. Jin Yamanaka, Shigesumi Kuwashima, and Takio Kurita. Fast and accurate image super resolution by deep cnn with skip connection and network in network. In *Neural Information Processing*, pages 217–225. Springer, 2017.
22. Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Transactions on graphics (TOG)*, 22(3):313–318, 2003.