Noname manuscript No.
(will be inserted by the editor)

M-FFN: Multi-Scale Feature Fusion Network for Image Captioning

Jeripothula Prudviraj · Chalavadi Vishnu · C. Krishna Mohan

Received: date / Accepted: date

Abstract In this work, we present a novel multi-scale feature fusion network (M-FFN) for image captioning task to incorporate discriminative features and scene contextual information of an image. We construct multi-scale feature fusion network by leveraging spatial transformation and multi-scale feature pyramid networks via feature fusion block to enrich spatial and global semantic information. In particular, we take advantage of multi-scale feature pyramid network to incorporate global contextual information by employing atrous convolutions on top layers of convolutional neural network (CNN). And, the spatial transformation network is exploited on early layers of CNN to remove intra-class variability caused by spatial transformations. Further, the feature fusion block integrates both global contextual information and spatial features to encode the visual information of an input image. Moreover, spatial-semantic attention module is incorporated to learn attentive contextual features to guide the captioning module. The efficacy of the proposed model is evaluated on the COCO dataset.

Keywords Image captioning · convolutional captioning · language attributes · language generation

1 Introduction

Automatically generating a natural language sentence by distilling the content of an image, often termed as image captioning, is a challenging task that connects both computer vision and natural language processing (NLP). This task goes beyond the conventional tasks such as classification [1] and object detection [2,3] as it requires a model to capture the holistic representation of an input image, and describe the content of visual scene in natural language.

Indian Institute of Technology Hyderabad.
Hyderabad, India.
E-mail: cs17resch01005@iith.ac.in

Captioning an image is an emerging challenge in visual scene understanding, and advocates the number of potential applications in the field of computer vision. It can aid visually impaired people, reinforce the content search on streaming platforms, strengthen the robotic vision, and allow users to organize and navigate unstructured visual data.

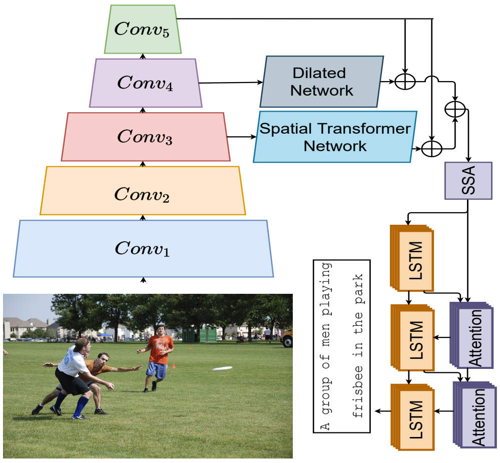


Fig. 1 Overview of the proposed multi-scale feature fusion network (M-FFN). Typically, the proposed M-FFN combines the spatial features of spatial transformer network, multi-scale semantic features of dilated network, and global features of ResNet backbone network to encode visual content of an image. The combined features are further fed to the attention based LSTM decoder to generate the caption of an image.

Image captioning has gained a lot of interest by bringing vision and language together. The prevalent approach to image captioning is an encoder-decoder framework [4–8], inspired from machine translation [9]. It explores the combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to model visual and textual features. The convolutional neural network encodes the visual information of an input image and then recurrent neural network (RNN) based decoder undertakes the encoded representation to generate a caption. Remarkable works on image captioning [10–12] learn a probabilistic model over the caption, conditioned on either image or object features.

Until recently, image captioning approaches leverage the visual information captured from the last convolutional layers of CNN [6,13] or the visual object region features extracted by the object detection module [11,14]. However, it is typically inadequate to use one visual feature to encode finer details of objects [15]. The potential drawbacks with the existing image captioning approaches are: i) Lack of fine-grained details and spatial context. ii) Suffer from high intra-class variability caused by spatial transformations. iii) Fail to incorporate multi-scale contextual information.

Recent methods [16,17] on visualizing the characteristics of each CNN layer depict that the features of early layers and top layers of CNN network are complementary by nature, where top layers of the network capture more semantic information but lack fine-grained details [18]. And, early layers features contribute spatial details of an image but lack semantic information [19]. Therefore, spatial context and fine-grained details of local features are need to be captured together with the semantic information of an image [15]. Motivated by the above observations, we propose a feature fusion network to bridge the semantic and spatial gap between the early-level and top-level feature maps. Mainly, the proposed fusion framework concatenates the semantic information with the spatial context of local features to capture scene contextual information of an image.

Although, the feature fusion networks are extremely powerful for any computer vision task, their performance may degrade adversely when the data is embodied by intra-class variability caused by spatial transformations such as affine or perspective [20]. Therefore, attaining the invariance to such transformations is highly beneficial for vision tasks such as image captioning. One of the effective methods to bring invariance to geometric transformations in deep convolutional network is the spatial transformer network (STN) [21]. Henceforth, we leverage the STN framework that can be seamlessly incorporated into our proposed feature fusion network to transform the input data and further achieve spatial invariance.

In addition, we capture the multi-scale contextual information to encode effective scene contextual information by exploring the semantic features with various filters at multiple dilation rates and multiple field-of-views. In particular, we apply several parallel atrous/ dilated convolutions on semantic features with different rates. Even though the top layers of CNN network holds rich semantic information, the definite information of object boundaries is lacking due to the recurrent pooling or convolutions with striding operations within the encoder network. This issue could be addressed by incorporating the atrous/ dilated convolution over the extracted semantic feature.

Attempting to combine the advantages of aforementioned methods in this work, we first point out the simple fusion network with semantic and spatial information for vision-to-language tasks to account for discriminative features along with the scene contextual information. Then, we exploit the spatial transformation network on early layers of convolutions to remove intra-class variability caused by spatial transformations. Further, we employ dilated network with several atrous convolutions on semantic feature map to extract contextual information of an image at multiple scales. In addition, we incorporate spatial-semantic attention (SSA) module to learn the contextual relations among the spatial and semantic features. The overview of the proposed multi-scale feature fusion network is illustrated in the Figure 1.

In a nutshell, the proposed multi-scale feature fusion network (M-FFN) builds on the observation that although the top layers of convolutional networks are more effective to capture semantic information, they are inadequate to encode the fine-grained details and spatial context such as small objects and spatial relationships. The earlier layers, on the other hand, are rich in spatial concepts but do not capture semantics. This observation provokes to construct multi-scale feature fusion network. Further, to achieve invariance and multi-scale information, we take advantage of spatial transformer network and dilated network. The main contributions of our work are summarized as follows.

- We propose a multi-scale feature fusion network to effectively encode the fine-grained details of local features and the scene contextual information of an image for image captioning task.
- We build spatial transformer network on early layers of fusion network to incorporate invariance for the spatial features and probe the dilated network to achieve multi-scale contextual information.
- A novel feature fusion block and spatial-semantic attention module is exploited to align various visual features which further directed to caption decoding module.
- The proposed feature fusion network outperforms the significant works of image captioning on COCO dataset.

2 Related work

In this section, we first present the various feature encoding mechanisms that are exploited semantic information, spatial concepts, multi-scale contextual information, and spatial invariant features on several computer vision tasks. Then, we review the prominent works of image captioning task.

2.1 Visual feature encoding networks

Farahzadeh *et al.* [15] proposed a framework to learn weighted combination of local semantic topics along with global and spatial information for scene action recognition. A combined coarse and fine semantics information is explored via shortcut connection fusion block in [18] to model feature correlation and align the extracted features of two different domain. Ma *et al.* [22] exploited hierarchical features of deep CNN to learn adaptive correlation filters on the outputs of each convolutional layer for robust visual tracking. Mishra [23] *et al.* introduced a novel fuzzy inferencing technique combined with classical CNN network to extract efficient frame features for action recognition. Recently, Ding [24] *et al. propose stimulus-driven attention and the concept-driven attention on CNN+LSTM architecture for image captioning task.* To bridge the gap between low-level and high-level features, Zhang *et al.* [19] proposed ExFuse network which introduces semantic information into spatial features and spatial details into high-level features. Si *et al.* [25] introduced a multi-features fusion module to obtain spatial features and enlarge receptive field. In [26], the three stage inference information processes are assembled to explicitly model the information flows and structures for human parsing task. Further, Lu *et* *al.* [27] proposed a CD-LinkNet network to combine the multi-level features in order to resolve deformation and multi-scale variations.

The spatial transformer network (STN) [21] is the well accepted method on improving the CNNs to remove spatial transformations such as affine or perspective. This learnable module in the network allows the spatial manipulation of data without any explicit supervision. Annunziata *et al.* [20] proposed dense fusion network by combining multiple STNs, namely, densely fused spatial transformer network (DeSTNet). Luo *et al* [28] incorporates spatial transformer network module in person re-identification task to sample an affined image from the holistic image in order to match the partial image. The oriented spatial transformer network (OSTN) is presented in [29] to detect the pedestrians in fish-eye images effectively. The multi-scale context information can be achieved by incorporating atrous convolutions in the CNN network. Chen *et al.* [30] utilized atrous convolutions for semantic image segmentation where atrous convolutions were used to explicitly control the resolution of Deep CNNs and enlarge the field of view of filters effectively in order to employ larger context without increasing the number of parameters. Further, it is combined in encoder-decoder framework [31] to probe the CNN features at multiple rates for encoding the multi-scale contextual information.

2.2 Image captioning

Current surge of research interest in captioning an image is outgrowing by bridging contextual visual representation and natural language expression. The prevalent approach for image captioning is encoder-decoder framework, where convolutional neural network is utilized as the encoder for semantic visual representation, while recurrent neural network (RNN) is often incorporated as the decoder for generating a caption. Vinyals *et al.* [32] presented the NIC model where the pre-trained CNN model is leveraged to extract image features and long-short term memory network is utilized to generate caption. Inspired by CNN-LSTM based framework, Jia *et al.* [33] introduced the gLSTM approach where global visual features are incorporated with rich semantic information to guide caption decoding module. A deep hierarchical encoder-decoder network is presented in [34] for image captioning, where it explores the vertical depth of encoder-decoder framework.

In the conventional CNN-LSTM framework, visual information may be lost or intruded by the visual noises while generating natural language sentence. To mitigate such issues, the attention mechanism offers more local visual clues to language module. In other words, the visual features extracted at the last convolutional layer of CNN are selectively focuses on prominent regions of an image while generating the sentence. Xu *et al.* [35] proposed soft and hard attention mechanisms for image captioning task. Similarly, Lu *et al.* [36] introduced a novel adaptive attention model with a visual sentinel to learn to focus on either image regions or to the visual sentinel at a given time-step. The spatial and channel-wise attention in convolutional layers is incorporated in [37] to attend spatial locations of feature maps at multiple layers and multiple channels. Anderson *et al.* [14] applied attention at the level of objects and other salient regions of an image by leveraging bottom-up and top-down attention mechanisms. By feeding attended image regions solely to the captioning module, the CaptionNet achieved impressive performance on image captioning task. Recently, Feng *et al.* [7] introduced a cascaded revision network for novel object captioning without external domain knowledge. Further, a random image cropping and patching technique is introduced in [38] to augment new images to the dataset for image captioning task.

Although the existing approaches of image captioning task are showing significant performance, they fail to encode attentive multi-scale contextual features of an image to generate caption of an image. Motivated by this, we propose a novel feature fusion process that learns multi-scale contextual information of an image and generates fine-grain captions.

3 Encoder-decoder model for image captioning

This section describes the widely accepted encoder-decoder framework [32] for image captioning. Let, an image I to be captioned by a sentence S, where S = $\{w_1, w_2, w_3, \ldots\}$ consisting of N_s words (w). We first encode the semantic information W_I of an image using deep CNN. Then, a long-short term memory (LSTM) network is employed to generate the captioning. At each time step t, the LSTM network recursively process the encoded information using input gate i_t, forget gatef_t, output gate o_t, and context vector c_t are given by

$$\begin{aligned} i_t &= \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i), \\ f_t &= \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f), \\ o_t &= \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o), \\ c_t &= i_t \odot \phi(W_{cx}x_t + W_{ch}h_{t-1} + b_c) + f_t \odot c_{t-1}, \\ h_t &= o_t \odot tanh(c_t), \\ z_t &= W_z h_t, \end{aligned}$$

where ϕ is maxout non-linearity and σ is the sigmoid activation function. W_*, V_*, U_*, and b_* are the parameters to be learned. The distribution over the next word for obtained h_t and c_t can be defined as

$$w_t = softmax(z_t). \tag{1}$$

For a given input image-caption pair, the objective of the encoder-decoder model is to minimize the loss on the model parameters θ of the model as

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(w_t|w_1, w_2, \ldots, w_{t-1}, I)), \tag{2}$$

where $(p_\theta(w_t|w_1, w_2, \ldots, w_{t-1}, I))$ is devised from Equation 1

4 Proposed multi-scale feature fusion network for image captioning

In this work, we exploit feature fusion problem in the encoder-decoder based image captioning framework. In general, the pre-trained convolutional neural network is employed as encoder module, where it generates spatial features from early layers and semantic features from top layers. The decoder part acquires the semantic features and then generates a natural language sentence using recurrent neural networks such as LSTMs. Farahzadeh *et al* [15] argue that the utilization of single visual feature vector is often insufficient to represent the scene contextual information. In other words, the semantic features extracted at top layers fail to encode finer details of small objects. To mitigate this problem, a feature fusion method is proposed in computer vision community for various tasks such as segmentation [19], human parsing [26], action recognition [39], object detection [2], and so on. Typically, the feature fusion network generates the low-level but high-resolution features from the bottom layers and high-level low-resolution features from the top layers and then combines the features from both top-down and bottom-up path way in order to achieve scene contextual information of an image. However, the spatial features of bottom layers adversely affected by the intra-class variability caused by the spatial transformations. And, the semantic features extracted from top-layers failed to incorporate the multi-scale contextual information of an image. Hence, we propose a novel multi-scale feature fusion network which address the aforementioned challenges.

In a nutshell, we present the encoder-decoder network for image captioning, as shown in Figure 2, where encoder network is constructed using novel multi-scale feature fusion network and decoder module contains LSTM based captioning framework. In the following sections, we present the proposed visual encoder module and caption decoding module.

4.1 Visual encoder

The visual encoder is a feature extractor network that extracts multi-scale scene contextual information of an image by processing given input image through various sub-network modules. Mainly, the proposed visual encoder network consists of five components, namely, backbone network, spatial transformer network, dilated network, feature fusion block, and spatial-semantic attention module.

4.1.1 Backbone network

We use ResNet [1] model pretrained on ImageNet [40] as the backbone feature extractor for our visual encoder in order to encode the visual representaion of an image. The ResNet network consists of 5 layers, namely, $Conv_1$, $Conv_2$, $Conv_3$, $Conv_4$, and $Conv_5$, where each $Conv$ layer contains varied number of Bottle-Neck modules. The semantic information of different objects

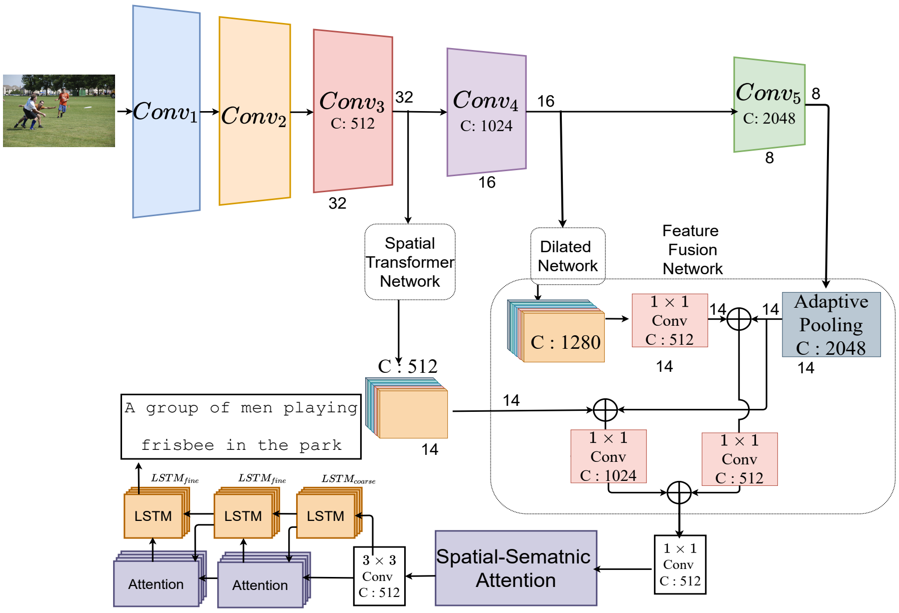


Fig. 2 Framework of the proposed multi-scale feature fusion (M-FFN) network. The proposed M-FFN has five major components, i.e., ResNet-101 backbone network, spatial transformer network, dilated network, feature fusion block, and spatial-semantic attention module. The ResNet backbone network outputs the global features from $Conv_5$ layer, semantic features from $Conv_4$ layer, and spatial features from $Conv_3$ layer. Then, the spatial transformer takes the spatial feature of $Conv_3$ and outputs the spatial invariant features. Further, The dilated network incorporates the multi-scale semantic features. A feature fusion block combines the spatial-invariant features, multi-scale semantic features, and global features to achieve multi-scale contextual information of an image. An attention layer selectively focuses on prominent features of multi-scale contextual features and feeds through LSTM caption decoding module. The LSTM network takes the attention weights, word embeddings and generates the precise and diverse captions.

is strengthened as visual features are propagated from bottom to top layers of network. And, the spatial information of objects in a visual scene gradually reduces. Henceforth, we propose to learn the deep network which combines the advantages of both spatial and semantic features of an image to generate a caption.

The overall image captioning framework follows the CNN-LSTM architecture, as shown in Figure 2. From the backbone ResNet framework, we first remove the fully-connected layers as they expose little spatial resolution of 1×1 and utilize only the hierarchical features of $Conv$ layers. The proposed Multi-sccale feature fusion network (M-FFN) extracts the 3 different semantic levels of feature maps from the ResNet model, namely, $Conv_3$, $Conv_4$, and $Conv_5$. Given the input image, the extracted $Conv$ layers produce the spatial resolutions of 32×32, 16×16, and 8×8 with the channel size of $512, 1024$, and 2048, respectively. From $Conv_3$ to $Conv_5$, the spatial resolutions is gradually decreasing and semantic information is increasing. Hence, we leverage dilation network on $Conv_4$ and spatial transformer network on $Conv_3$ layers to bridge the gap between low-level and high-level featues. Further, we introduce feature fusion block to embed the effective visual features. In the following sections, we will elaborate our spatial transformer network, dilated network, and feature fusion block in detail.

4.1.2 Spatial transformer network

The convolutional backbone networks are exceptionally prevailing models for image captioning, but their performance is limited by the lack of ability to handle the spatial invariance of the input data. The utilization of local max-pooling operations has aided to the network by allowing it to learn the spatial invariance of the positional features. However, this spatial invariance is only achieved through a deep network of multiple max-pool and convolutional layers. And, the intermediate feature maps like $Conv_2$ and $Conv_3$ of CNN are not truly invariant to large transformations of the input data [21,41]. This limitation of CNNs is because of having only a confined and pre-constrained polling operations for dealing with spatial transformations of data.

Unlike max-pooling layers, which are fixed and local to receptive fields, the spatial transformer network (STN) is the appropriate method of choice for CNNs to give the ability to remove the spatial transformations and improves performance in an end-to-end network. The STN modules are dynamic and spatially transform the feature map by generating a suitable transformation for each input image. The transformation applied on each feature map handles rotations, cropping, scaling, as well as non-rigid deformations. This not only allows spatial transformations in the network but also detects the regions of an image that are most important [21].

Motivated by the above observations, we employ spatial transformer module on $Conv_3$ feature map of ResNet backbone network to allow the spatial invariance with in the network. This network actively transforms the features spatially by conditioning on itself without any additional training supervision and modification to the optimization method. As shown in Figure 3, the spatial transformer network contains three components, namely, the localization network, the grid generator, and the sampler. First, the localization network takes the input feature map and outputs the parameters of spatial transformations that are to be applied on the feature map. Then the grid generator takes the predicted transformation parameters to create a sample grid. The sample grid consists of set of points which are sampled from the transformed output. Finally, the sampler takes feature map and sampling grid as inputs and produces the output the sampled map of input and grid points. The combination of these three components is called spatial transformer network and the functioning of each component is described as follows.

Localization network: Given input feature map $F \in \mathbb{R}^{W \times H \times C}$ with width W, height H, and channels C to the localization network, it applies the transformation $\mathcal{T}_\theta$ on the feature map and outputs the transformation parameters θ, $\theta = Trans_{loc}(F)$. Here, the size of θ depends on the transformation type

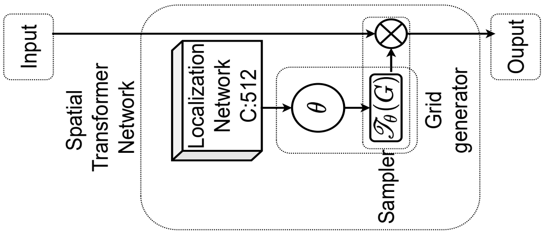


Fig. 3 Spatial transformer network

such as affine or perspective. In our work, we utilize affine transformations. Hence, the size of θ is a 6- dimensional as in Equation 3.

Grid generator: The grid generator takes the output of localization network and applies a sampling kernel to wrap the input feature map. For instance, consider $\mathcal{T}_\theta$ is a 2D affine transformation Q_θ such as image and then the output pixels are defined to lie on a regular grid $Z = Z_i$ of pixels $Z_i = (x_i^t; y_i^t)$ is formulated using

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \mathcal{T}_\theta(Z_i) = Q_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} \; \theta_{12} \; \theta_{13} \\ \theta_{21} \; \theta_{22} \; \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}, \tag{3}$$

where $(x_i^t; y_i^t)$ and $(x_i^s; y_i^s)$ are the target coordinates of output feature map and source coordinates in the input feature map, and Q_θ is the affine transformation matrix.

Sampler: The sampler takes the set of sampling points $T_{(G)}$, along with the input feature map F to perform spatial transformation of the feature map and produces the sampled output feature map O_f using

$$V_i^c = \sum_n^H \sum_m^W F_{nm}^c g(x_i^s - m; \phi_x) g(y_i^s - n; \phi_y)$$
$$\forall i \in [1 \ldots H'W'] \, \forall c \in [1 \ldots C] \, . \tag{4}$$

In brief, the spatial transformer network formed by the combination of localization network, grid generator, and sampler allows our encoder network to learn to actively transform the convolutional feature maps and thus provides spatial invariance to the early layers like $Conv_3$ in our encoder network.

4.1.3 Dilated network

Deep CNNs are showing great success on computer vision tasks such as image classification, object detection, and high-level tasks such as image captioning by learning semantic features. Although the top convolutional layers of CNN backbone network have rich semantic information, the finer details of object boundaries are missing due to the repeated pooling and strided convolutions with in the backbone network. To address this issue, we employ dilated network

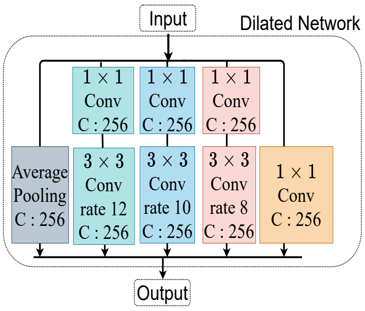


Fig. 4 Dilated network

(Figure 4) on $Conv_4$ layer of ResNet backbone network to extract denser feature information.

We construct dilated network with several parallel dilated convolutions to capture the scene contextual information at multiple scales. In particular, dilated convolutions allow us to explicitly control the resolutions of feature maps and helps to enlarge the field of view of filters which further employ larger context information without increasing the number of parameters and amount of computation. In addition, the parallel dilated convolutions in the dilated network robustly segment the objects at multiple scales and probes each layer with filters at multiple sampling rates and effective fields-of-views, thus it incorporates multi-scale semantic information. Given a two dimensional signal such as image, for each spatial location p on the output feature map k and a convolutional filter w, the dilated convolution is applied over the input feature map j as follows

$$k[p] = \sum_{l} j[p + r \cdot l]w[l], \tag{5}$$

where r denotes the dilation rate at which we sample the input and $r = 1$ is a special case that demonstrates the standard convolution. With the change of dilation rate, the filter's field of view changes adaptively.

4.1.4 Feature fusion block

The purpose of feature fusion block is to achieve the multi-scale scene contextual information of an image by combining the spatial features, multi-scale semantic features, and global features. In the proposed framework, we extract spatial features from low-level layers of ResNet backbone network i.e., $Conv_3$. Usually, the features extracted from early layers of deep CNN network are not spatially invariant to the input data. Hence, we employ spatial transformer network to make extracted $Conv_3$ features spatially invariant. Typically, the spatial generator network learns to limit the intra-class variance caused by spatial transformations and produces the feature map which is equal to the input feature map. Thus, we achieve spatial features that are freed from affine transformations.

On the other hand, the high level features extracted from $Conv_4$ layer of the backbone network are limited with spatial details. Thus, it looses detailed information of objects. To encode the finer details of objects, we leverage dilated network which is able to control the feature resolution without retraining the backbone network. Further, we employ parallel atrous convolutions to encode the semantic information at multiple scales.

In addition to spatial and multi-scale semantic features, the global semantic features is extracted from the top layer of backbone network i.e., $Conv_5$. Since the spatial size of top-layer feature maps ($Conv_4$ and $Conv_5$) is very small, we adaptively pool the all features to spatial size of $Conv_3$ feature map. The extracted feature maps of $Conv_3, Conv_4$, and $Conv_5$ are further concatenated and processed with standard 1×1 convolutions before feeding it to attention module.

On extracting convolutional features C_3, C_4, and C_5 from $Conv_3, Conv_4$, and $Conv_5$, respectively, the mathematical formulation of the feature fusion block can be defined as:

$$\begin{aligned}
B_1 &= STN(C_3) \\
B_2 &= DN(C_4) \\
Fusion_1 &= Conv(B_1) + pool(C_5) \\
Fusion_2 &= Conv(B_2) + pool(C_5) \\
Feature\_fusion &= Conv(Fusion_1) + Conv(Fusion_2),
\end{aligned} \tag{6}$$

where B_1 and B_2 represent the output features of spatial transformer network (STN) and dilated network (DN). $Conv$ denotes the convolutional operation.

In brief, the visual encoder is incorporated with rich spatial, semantic, and global features. Thus, it effectively represents the multi-scale contextual information of an image. Further, we elaborate the caption decoding module with the context of encoded feature representation.

4.1.5 Spatial-semantic attention

Recently, the self-attention mechanisms are widely explored for various computer vision tasks [42,43] to incorporate long range dependencies among visual features. Inspired by its effectiveness on context modelling, we leverage self-attention in our proposed approach to capture the contextual relations among spatial features and semantic features.

Given a concatenated spatial and semantic features M of size $H \times W \times C$, the spatial-semantic attention (SSA) module constructs an attention map A of size $C \times C$ to learn the contextual relational features. As shown in Figure 5, given an input $M \in \mathbb{R}^{N \times C}$, we first generate two feature maps B and E of size $N \times C'$ by employing convolutional layers, where $N = H \times W$ and $C' = C/\tau$. Then, we perform matrix multiplication between transpose of E and B, resulting in a $C' \times C'$ attention map A. The obtained attention map A is multiplied with convolutional feature map $J \in \mathbb{R}^{N \times C'}$ and fed to the another convolutional layer to restore the same size of M. The resulting feature

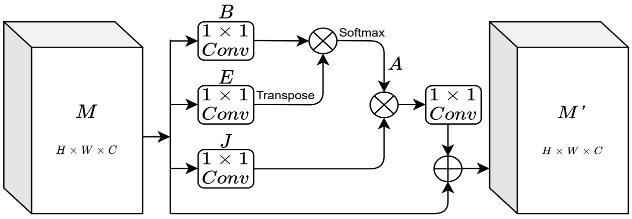


Fig. 5 Spatial-semantic attention

map is concatenated with the input feature map M via residual connection to generate the output feature map $M' \in \mathbb{R}^{N \times C}$. The computation of SSA can be formulated as

$$M' = W * (J \cdot A) + M, \tag{7}$$

where W denotes the convolutional kernel after A, and τ controls the output channel dimension. We experimentally set τ as 2.

4.2 Caption decoder

On achieving encoded representation of an image, we employ the attention mechanism before feeding it to the LSTM network in order to selectively choose the significant regions of visual encoder. Given a set of encoder features refereed as annotation vectors, the LSTM decoder learns to focus on significant regions of an image by selecting a subset of all the annotation vectors. Let $a = \{a_1, a_2, \ldots, a_L\}$ be the set of annotation vectors. Then, a_i, $i = \{1, 2, \ldots, L\}$ is the extracted feature vector at image region i. For each region i, the attention mechanism assigns a positive weight of α_i. We compute the attention weight α_i by designing the attention model q_{att} using the annotation vector a_i and previous hidden state of LSTM cell h_{t-1}. The attention module at image location i is formulated as

$$b_{ti} = q_{att}(a_i, h_{t-1}),$$
$$\alpha_{ti} = \frac{\exp(b_t i)}{\sum_{k=1}^{L} \exp e_{tk}}.$$

On obtaining the attention weights, the context vector (c_t) is updated as

$$c_t = \psi(\{a_i\}, \{\alpha_i\}), \tag{8}$$

where ψ is a function that outputs a single vector from the set of annotation vectors and their corresponding attention weights. Further, an average of the annotation vectors predicted by the hidden state of the LSTM and initial memory are fed through two separate multi-layer perceptrons (MLPs).

$$c_0 = g_{init,c}\left(\frac{1}{L}\sum_{i}^{L} a_i\right) \quad h_0 = g_{init,h}\left(\frac{1}{L}\sum_{i}^{L} a_i\right)$$

We generate the word probabilities given the context vector, LSTM hidden state, and previous words using

$$p(w_t) = -\sum_{t=1}^{T} \log(p_\theta(w_t|w_1, w_2, \ldots, w_{t-1}, a)), \tag{9}$$

where θ are the parameters of the model.

In this work, we leverage multi-stage LSTM network to generate rich fine-grained captions of an image. Typically, the multi-stage LSTM network is incorporated with one coarse LSTM ($LSTM_{coarse}$) and two finer LSTM networks ($LSTM_{fine}$), where each LSTM network operates on the outputs of previous LSTM network and attention vector. In particular, the $LSTM_{coarse}$ generates the coarse-grained captions, and successive $LSTM_{fine}$ networks refine the generated captions and produce fine-grained captions. At each stage, we input previous LSTM hidden vector and attention weights.

5 Experimental results

In this section, we validate the performance of the proposed multi-scale feature fusion network (M-FFN) with quantitative and qualitative analysis. First, we present the quantitative analysis with state-of-the-art models using standard evaluation metrics like BLEU-n, METEOR, ROUGE L, and CIDEr. Then, the generated captions and their attention maps are illustrated to analyze the performance of our M-FFN qualitatively.

5.1 Dataset

We demonstrate the performance of the proposed M-FFN model by conducting experiments on two large datasets of image captioning, i.e., Flickr30K and COCO. Flickr30k dataset consists of 31k images with 158k crowd-sourced captions. The images of Flickr30k depict various events and activities performed by humans. Due to the lack of official split, We adopt publicly available split from [36, 44], which includes 29k images for training, 1000 images for both validation and test.

The COCO dataset provides $82k$ images for training, and $40k$ images for both validation and test set. This dataset renowned to be the challenging dataset as it contains multiple objects in the context of complex visual scenes. Each image in the dataset has annotated with 5 captions. To validate the model, we use widely adopted data split, which is $5K$ images for validation and test set, each. Further, we evaluate generated captions using conventional evaluation metrics like BLEU-1 and BLEU-4 [45], METEOR (MR) [46], ROUGE_L (RL) [47], and CIDEr (Cr) [48]. In brief, all these metrics compute the coherence between the n-gram occurrence in reference caption and generated caption. The coherence is weighted by the diversity and saliency of the ground-truth caption.

5.2 Implementation details

We implement the proposed M-FFN model using Pytorch framework. We utilize ADAM optimizer [49] with a learning rate of 0.0001 and 0.0004 for encoder and decoder. The batch size is set to 32 and learnt until the accuracy on the validation set does not change for 15 epochs. In addition, the decay rate is employed when the model does not improve for six consecutive epochs. The embedding dimension of visual encoder and attention layer is set to 2048 and 512, whereas the LSTM hidden layer and context layer dimension set to 512. And, we set the channel dimension of spatial transformer network to 512, dilated network to 256, and feature fusion network to 1024. In the following subsections, we will present the detailed implementation details of visual encoder and caption decoder modules.

5.2.1 Visual encoder

We use ResNet-101 network pre-trained on ImageNet as a backbone network for our visual encoder, where we extract convolutional features of size $32 \times 32 \times 512$, $16 \times 16 \times 1024$, and $8 \times 8 \times 2048$ from $Conv_3$,$Conv_4$, and $Conv_5$ layers, respectively. At first, we apply spatial transformer network on $Conv_3$ layer to incorporate spatial invariance to extracted spatial features. We employ spatial transformer network with *localization, regressor,* and *postconv* layers. The localization network is constructed using 5×5 filter with 512 channels and Max-pool with stride 2. As illustrated in Equation 3, we employ 3×2 affine matrix in regression network. For *postconv* layer, we use 5×5 filters with 512 channels. The output of *postconv* layer is concatenated with the adaptively pooled $Conv_5$ features to achieve spatial-semantic features of an image.

To achieve multi-scale semantic features, we incorporate dilated network on $Conv_4$ layer of ResNet backbone network. Typically, we employ parallel dilated convolutions with the filter size of 3×3 and dilation rates of $2, 4,$ and 6. In addition, we add average pooling and 1×1 convolutions to employ features from various filed of views. We set number of channels to 256 in the entire dilated network and concatenated all 5 layers features before feeding it to the adaptive pooling layer. The output features of adaptive pooling layer are concatenated with adaptively pooled $Conv_5$ features to encode rich multi-scale semantic features. Further, we concatenate the spatial features and multi-scale semantic features to produce multi-scale scene contextual information of an image.

5.2.2 Caption Decoder

At first, the attention layer takes the multi-scale contextual features and generates the attention weights of the feature map to selectively focus the important features of input feature. Then, the attention weights are fed to the LSTM network to generate the caption of an image. We utilize the standard LSTM cells to learn the word sequences using annotation vectors obtained from attention layer and input word embeddings at time instance t. The dimension of attention layer, LSTM hidden units, and LSTM context vector is fixed to 512 dimension.

5.3 Quantitative results

In this work, we present multi-scale scene contextual information for image captioning task using novel feature fusion network. The proposed multi-scale feature fusion network generates multi-scale contextual features of an image which are spatially invariant and semantically rich. Table 1 presents the results on Flickr30k dataset. From table 1, we can infer that the proposed model is outperforming the existing state-of-the-art models in all metrics. Notably, the proposed multi-scale feature fusion network($M - FFN_{dilated+STN+SSA}$) improves the state-of-the art on BLEU-4 from 35.8 to 39.5, and METEOR from 27.8 to 29.2, and CIDEr from 57.4 to 62.1. Further, Table 2 demonstrates the performance comparison with the other state-of-the-art models on the COCO dataset.

Table 1 Performance comparison on Flickr30k. B-n, M, R, and C stand for BLEU, Meteor, Rouge-L, and CIDEr, respectively

Model	**B-1**	**B-2**	**B-3**	**B-4**	**M**	**C**
DHEDN3 [34]	65.3	46.7	32.9	23.1	19.2	-
SCA-CNN [37]	66.2	46.8	32.5	22.3	19.5	-
SDCD [24]	66.3	43.7	29.2	21.1	-	-
Adaptive [36]	67.7	49.4	35.4	25.1	20.4	53.1
SCN+LSTM [44]	73.5	53.0	37.7	26.5	21.8	-
ARL [50]	75.9	60.3	46.5	35.8	27.8	57.4
$\mathbf{M - FFN_{dilated+STN}}$	76.1	60.7	51.4	37.3	27.6	59.1
$\mathbf{M - FFN_{dilated+STN+SSA}}$	**76.4**	**62.4**	**53.1**	**39.5**	**29.2**	**62.1**

On COCO dataset, the proposed M-FFN network is outperforming state-of-the-art models in all metrics. Mainly, we compare the performance of proposed models with recent encoder-decoder frameworks like deep hierarchical encoder-decoder framework (DHEDN) [34], compact image captioning model (COMIC) [53], salience-enhanced images (VIS-SAS) [59], transformer based captioning models [8,56,8]. In addition, we also included several attention based and attribute based models like spatial and channel-wise attention (SCA-CNN) [37], global-local attention (GLA) [52], semantic compositional networks (SCN) [44].

Moreover, the Table 3 demonstrates the significance of each component of the proposed approach on COCO captioning dataset. From Table 3, we can infer that the combination of dilated, STN, and SSA model produces the significant performance on the COCO dataset. In particular, our proposed

Table 2 Performance of the proposed models and other state-of-the-art methods on the COCO dataset

Model	BLEU-1	BLEU-4	METEOR	ROUGE_L	CIDEr-D
Deep VS [51]	62.5	23.0	19.5	-	-
NIC [32]	66.6	24.6	-	-	-
S-ATT [35]	70.7	24.3	23.9	-	-
SCA-CNN [37]	71.9	31.1	25.0	-	-
SCN [44]	72.8	33.0	25.7	-	101.2
GLA [52]	72.5	31.2	24.9	53.3	96.4
COMIC [53]	72.9	32.8	-	-	100.1
DHEDN3 [34]	73.1	32.3	25.6	53.7	99.3
Recall [54]	75.8	33.06	24.6	-	103.7
AOA [55]	80.2	38.9	29.2	58.8	129.8
M^2T [56]	80.8	39.1	29.2	58.6	131.2
MT_{umv} [8]	81.9	40.7	29.5	59.7	134.1
$OSCAR$ [57]	-	41.7	30.6	-	140.0
VinVL [58]	-	41.0	31.1	-	140.9
$\mathbf{M-FFN_{dilated+STN+SSA}}$	**84.1**	**41.9**	**31.6**	**64.1**	**141.4**

approach with spatial transformer network (STN) alone boosts the performance of captioning on Cider-D metric from 83.4 to 118.5 over the baseline CNN+LSTM approach [35]. Further, the dilated network, dilated network + spatial-semantic attention (SSA), and dilated network + STN report the performance of 120.9, 129.7, and 136.2, respectively, and showing better performance than the recent state-of-the-art approaches like Recall [54], AOA [55], and M^2T [56]. Finally, the combined model of dilated, STN, and SSA (M-FFN_dilated+STN+SSA) demonstrates the superior performance over the all existing state-of-the-art model in all standard captioning metrics.

Table 3 Performance of the different components on the COCO dataset

Model	BLEU-1	BLEU-4	METEOR	ROUGE_L	CIDEr-D
Baseline [35]	71.8	25.0	23.0	50.1	83.4
$\mathbf{M-FFN_{STN}}$	75.3	32.8	25.3	52.1	118.5
$\mathbf{M-FFN_{dilated}}$	75.7	33.1	25.6	53.4	120.9
$\mathbf{M-FFN_{dilated+SSA}}$	79.4	38.1	28.6	57.9	129.7
$\mathbf{M-FFN_{dilated+STN}}$	81.7	39.3	29.1	60.8	136.2
$\mathbf{M-FFN_{dilated+STN+SSA}}$	**84.1**	**41.9**	**31.6**	**64.1**	**141.4**

In brief, various encoding mechanisms have been proposed over the years to generate the effective caption of an image. For instance, the semantic information of an input image is encoded in the NIC [32] model to generate natural language sentence. Further, the soft and hard attention mechanism based encoder framework is proposed in [35] to selectively focus prominent regions of an image. Gan *et al.* [44] introduced a tag dependent weight matrix to decoding LSTM network to describe the content of an image. Recently, Linghui *et al.* [52] presented global and local attention to account for image-level and object-level features. Xinyu *et al.* [34] incorporated a hierarchical three-layer LSTM network to combine visual and textual concepts in order to generate captions.

In contrast to the existing works, we propose a novel encoding method that combines spatially invariant features and multi-scale semantic features through a feature fusion network. In particular, we employ spatial features to encode the finer details of small objects and multi-scale features are incorporated to achieve information of visual scene from various filed of views. However, the spatial details are sensitive to the spatial transformations such as affine. Hence, we employ spatial transformer network on spatial features $Conv_3$ and parallel dilation convolutions are incorporated to extract multi-scale semantic features.

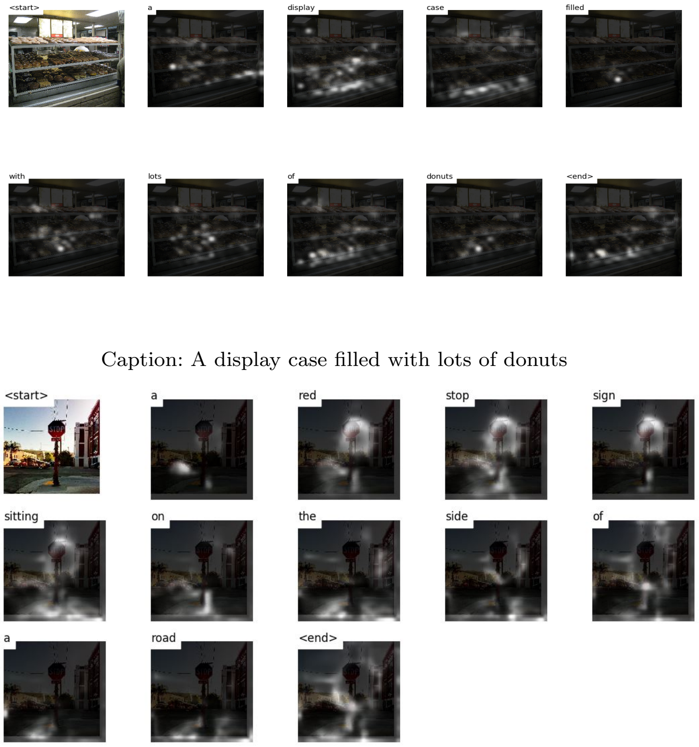
5.4 Qualitative results

In this section, we present the qualitative analysis of the proposed multi-scale feature fusion network (M-FFN) through the generated captions and attention maps of test set images. The generated captions and their attention maps are shown in Figure 6. From the Figure, we can observe that the attention mechanism is learning to focus on both coarse-grain and fine-grain details of encoded multi-scale contextual information by generating larger blobs for coarse-grain information and smaller blobs for fine-grain details. Also, we can examine that the generated captions are precise and diverse. In particular, the generated words "stop sign", "display case", and "donuts" indicate that the model is accounting for scene contextual information. And, the words like "red" and "side of the road" validate the semantic nature of the captions.

From the quantitative and qualitative results, we can observe that the extracted multi-scale contextual features encode the effective scene contextual information for an image and generate the precise and diverse captions. Although the proposed model is outperforming the recent works of image captioning, we can improve our method by probing various attention mechanisms. Further, we can also take advantage of semantic attributes as in [44], visual relations, and object regions [14].

6 Conclusion

Recent works on image captioning leverage either attention mechansim with object features [14, 52] or visual attributes [44] with semantic information for describing the content of an image. However, these methods are limited with spatial and multi-scale semantic information of the visual scene. To incorporate multi-scale scene contextual information of an image, we propose a multi-scale feature fusion (M-FFN) network for image captioning. The proposed multi-scale feature fusion network actively learns the global features from top layer ($Conv_5$) of ResNet backbone network and multi-scale semantic features from the dilated network employed on semantic layer ($Conv_4$). In addition, we incorporate spatial features along with multi-scale semantic features to account for finer details of an image. However, the spatial features are ineffective to encode rich spatial information due to the intra-class variability caused by spatial



Caption: A display case filled with lots of donuts

Caption: a red stop sign sitting on the side of a road

Fig. 6 Illustration of attention maps and generated captions of test images.

transformations. To mitigate this problem, we employ a spatial transformer network on spatial ($Conv_3$) features. By combining spatial invariance features and multi-scale semantic features, the proposed model is able to encode the global, local, and spatial features of an image. Further, the spatial-semantic attention mechanism allows the caption decoding model to selectively focus on prominent features of scene contextual information. The effectiveness of the proposed model is demonstrated on COCO dataset by qualitatively and quantitatively.

References

1. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

2. Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
3. Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2896–2907, 2017.
4. Xinwei He, Yang Yang, Baoguang Shi, and Xiang Bai. Vd-san: Visual-densely semantic attention network for image caption generation. *Neurocomputing*, 328:48–55, 2019.
5. Jinsong Su, Jialong Tang, Ziyao Lu, Xianpei Han, and Haiying Zhang. A neural image captioning model with caption-to-images semantic constructor. *Neurocomputing*, 367:144–151, 2019.
6. Fen Xiao, Xue Gong, Yiming Zhang, Yanqing Shen, Jun Li, and Xieping Gao. Daa: Dual lstms with adaptive attention for image captioning. *Neurocomputing*, 364:322–329, 2019.
7. Qianyu Feng, Yu Wu, Hehe Fan, Chenggang Yan, Mingliang Xu, and Yi Yang. Cascaded revision network for novel object captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3413–3421, 2020.
8. Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multiview visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12):4467–4480, 2019.
9. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
10. Ying Hua Tan and Chee Seng Chan. Phrase-based image caption generator with hierarchical lstm network. *Neurocomputing*, 333:86–100, 2019.
11. Xiaodan Zhang, Shengfeng He, Xinhang Song, Rynson WH Lau, Jianbin Jiao, and Qixiang Ye. Image captioning via semantic element embedding. *Neurocomputing*, 395:212–221, 2020.
12. Dexin Zhao, Zhi Chang, and Shutao Guo. A multimodal fusion approach for image captioning. *Neurocomputing*, 329:476–485, 2019.
13. Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. Mscap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4204–4213, 2019.
14. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.
15. Elahe Farahzadeh, Tat-Jen Cham, and Wanqing Li. Semantic and spatial content fusion for scene recognition. In *New Development in Robot Vision*, pages 33–53. Springer, 2015.
16. Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks. In *Bernstein Conference 2015*, pages 219–219, 2015.
17. Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.
18. Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE international conference on computer vision*, pages 5551–5560, 2017.
19. Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–284, 2018.
20. Roberto Annunziata, Christos Sagonas, and Jacques Calì. Destnet: Densely fused spatial transformer networks. *arXiv preprint arXiv:1807.04050*, 2018.
21. Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.
22. Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang. Robust visual tracking via hierarchical convolutional features. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2709–2723, 2018.

23. Soumya Ranjan Mishra, Tusar Kanti Mishra, Goutam Sanyal, Anirban Sarkar, and Suresh Chandra Satapathy. Real time human action recognition using triggered frame extraction and a typical cnn heuristic. *Pattern Recognition Letters*, 135:329–336, 2020.
24. Songtao Ding, Shiru Qu, Yuling Xi, and Shaohua Wan. Stimulus-driven and concept-driven analysis for image caption generation. *Neurocomputing*, 398:520–530, 2020.
25. Haiyang Si, Zhiqiang Zhang, Feifan Lv, Gang Yu, and Feng Lu. Real-time semantic segmentation via multiply spatial fusion network. *arXiv preprint arXiv:1911.07217*, 2019.
26. Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5703–5713, 2019.
27. Yu Lu, Muyan Feng, Ming Wu, and Chuang Zhang. C-dlinknet: considering multi-level semantic features for human parsing. *arXiv preprint arXiv:2001.11690*, 2020.
28. Hao Luo, Wei Jiang, Xing Fan, and Chi Zhang. Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. *IEEE Transactions on Multimedia*, 22(11):2905–2913, 2020.
29. Yeqiang Qian, Ming Yang, Xu Zhao, Chunxiang Wang, and Bing Wang. Oriented spatial transformer network for pedestrian detection using fish-eye camera. *IEEE Transactions on Multimedia*, 22(2):421–431, 2019.
30. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
31. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
32. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.
33. Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE international conference on computer vision*, pages 2407–2415, 2015.
34. Xinyu Xiao, Lingfeng Wang, Kun Ding, Shiming Xiang, and Chunhong Pan. Deep hierarchical encoder–decoder network for image captioning. *IEEE Transactions on Multimedia*, 21(11):2942–2956, 2019.
35. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.
36. Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017.
37. Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017.
38. Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Data augmentation using random image cropping and patching for deep cnns. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2917–2931, 2019.
39. Zhigang Tu, Wei Xie, Justin Dauwels, Baoxin Li, and Junsong Yuan. Semantic cues enhanced multimodality multistream cnn for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(5):1423–1437, 2018.
40. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
41. Taco S Cohen and Max Welling. Transformation properties of learned visual representations. *arXiv preprint arXiv:1412.7659*, 2014.

42. Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.
43. Jean-Paul Ainam, Ke Qin, and Guisong Liu. Self attention grid for person reidentification. *arXiv preprint arXiv:1809.08556*, 2018.
44. Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5630–5639, 2017.
45. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
46. Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
47. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
48. Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
49. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
50. Junbo Wang, Wei Wang, Liang Wang, Zhiyong Wang, David Dagan Feng, and Tieniu Tan. Learning visual relationship and context-aware attention for image captioning. *Pattern Recognition*, 98:107075, 2020.
51. Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
52. Linghui Li, Sheng Tang, Yongdong Zhang, Lixi Deng, and Qi Tian. Gla: Global–local attention for image description. *IEEE Transactions on Multimedia*, 20(3):726–737, 2017.
53. Jia Huei Tan, Chee Seng Chan, and Joon Huang Chuah. Comic: Toward a compact image captioning model with attention. *IEEE Transactions on Multimedia*, 21(10):2686–2696, 2019.
54. Lingxiang Wu, Min Xu, Jinqiao Wang, and Stuart Perry. Recall what you see continually using gridlstm in image captioning. *IEEE Transactions on Multimedia*, 22(3):808–818, 2019.
55. Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643, 2019.
56. Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020.
57. Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.
58. Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.
59. Lian Zhou, Yuejie Zhang, Yu-Gang Jiang, Tao Zhang, and Weiguo Fan. Re-caption: Saliency-enhanced image captioning through two-phase learning. *IEEE Transactions on Image Processing*, 29:694–709, 2019.