Enhancing Construction Image Captioning with Dual Augmentation Methods: Synonymous Replacement and Contextualised Word Embeddings

Haosen Chen, Lei Hou*, Shaoze Wu, Guomin (Kevin) Zhang
RMIT University, Australia
lei.hou@rmit.edu.au

Abstract. In the construction industry, image captioning plays a crucial role in facilitating communication, documentation, and monitoring of construction projects. However, limited data availability and diversity present challenges to the development of accurate and diverse construction image captioning models. In this paper, we employ two augmentation methods, contextualised word embedding and synonymous replacement, to enhance the performance of image captioning models in the construction domain. An ablation study was conducted using a deep learning model to assess the effectiveness of the proposed methods. The results demonstrated that both augmentation methods individually and combined improved the model performance across all evaluation metrics, including BLEU, METEOR, ROUGE-L, CIDEr, and SPICE, with the combined method yielding the highest improvement. This research contributes to the construction safety monitoring and analysis field by providing an effective strategy for enhancing construction image captioning models' accuracy and diversity, ultimately improving project outcomes and overall efficiency.

1. Introduction

The construction industry is characterised by complex, dynamic, and time-sensitive projects that require effective communication and documentation to ensure success (Chen et al., 2023). In recent years, advances in digital imaging technologies have facilitated the capture and analysis of visual data from construction sites, providing valuable insights for project management, progress monitoring, and quality control (Hou et al., 2021; Li et al., 2020; Moon et al., 2022; Son and Kim, 2021; Xu et al., 2021). Although vision-based construction safety monitoring has been drawn considerable attention, it is still in a nascent stage. Benefited from recent advances in Deep Learning (DL), image captioning opens up a new avenue for construction safety monitoring and analysis. Image captioning, which involves the automatic generation of textual descriptions for images, can further enhance the utility of these visual data by providing contextually relevant and human-readable information. The image captioning model learns inter-modal correspondence between textual captions and visual features using a dataset of images with various descriptions. The structured text together with as-is onsite images can facilitate construction safety inspection and help in decision-making.

Although the algorithmic optimisation of image captioning models has been improved to some extent (Liu et al., 2020; Wang et al., 2022), the development of accurate and diverse construction image captioning models is often hindered by the limited availability of annotated data. Also, public datasets such as MSCOCO (Vinyals et al., 2016) and Flickr30K (Plummer et al., 2015) do not cover a wide range of construction scenarios. DL models trained on these datasets may misinterpret construction scenes. As a result, many models have low performance in terms of the varieties of descriptions generated. In specialised domains like construction, obtaining a large and diverse dataset of labeled images with corresponding captions can be challenging due to the need for domain-specific knowledge, time-consuming annotation processes, and the dynamic nature of construction projects. Consequently, the performance and generalisation of image captioning models in construction applications may suffer from data scarcity and limited diversity in training data.

To address these challenges, this study aims to bridge the research gap by proposing text augmentation methods that leverage domain-specific lexical substitution and contextualised word embedding to expand and diversify the dataset of construction image captions. By constructing a tailored thesaurus for synonymous replacements and employing state-of-the-art language models to generate contextually appropriate alternative captions, this method aims to improve the performance of construction image captioning models and better cater to the industry's specific language requirements. The resulting augmented dataset is expected to facilitate more effective communication, documentation, and monitoring within construction projects, ultimately contributing to improved project outcomes and overall efficiency.

2. Methodology

2.1 DL-based image captioning module

This study employed the image captioning model developed by Chen et al. (2021b). The network is composed of nodes that represent items and edges that represent connections between object groups. First, the Transformer model encodes the region of interest identified by Faster R-CNN. A scene graph is then constructed using edges and nodes corresponding to the identified regions of interest, and the graph representation is subsequently enriched with a graph convolutional network. The learnt semantic matrix is then supplied into the attention-based fusion module, allowing the model to process both semantic linkages and visual information. The structure of the image captioning model is demonstrated in Fig. 1.

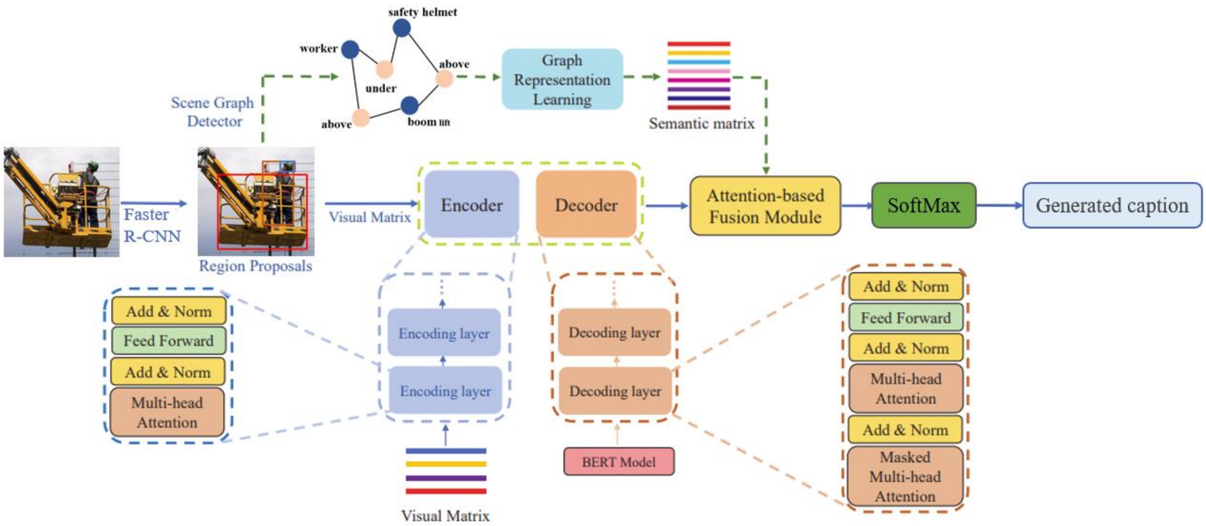

Figure 1. Architecture of the DL-based image captioning model.

In order to better capture the interrelationships between the visual regions of the image, a transformer encoder consisting of N identical coding layers is adopted. The input image is reformed to a set of visual feature vectors $V = [v_1, v_2, v_3, \dots, v_n]$. In order to prevent the loss of global visual information during convolution process, a global feature extraction is conducted. By doing this, the input image can be depicted as a visual matrix representation, which can be directly encoded via encoding layers.

In each multi-head attention layer H_i, it takes the visual matrix X in the form of three parameters, namely, Query (Q), Key (K), and Value (V) by multiplying three trainable weight matrix W^Q, W^K, W^V respectively. The Attention module repeats its computations multiple times in parallel, which is called Multi-head. and then the attention-based fusion module is employed as:

$$Attention\ (Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

$$MultiHead(Q.K.V) = Concat(H_1, \dots, H_h)W^0 \tag{2}$$

$$H_i = Attention\ (XW_i^Q, XW_i^k, XW_i^V) \tag{3}$$

Visual areas (represented by nodes) and their connections (represented by edges) are denoted on the scene graph. The BERT model is utilised for converting each node into tokens. Tokens are fed into BERT using Word-piece embeddings. Each token in BERT is represented by an embedding made up of a token embedding, a location embedding, and a segment embedding. Token ordering information is stored in positional embeddings. Consequently, the graph can be represented as node feature matrix $X = [x_1, x_2, x_3, \dots, x_n]^T$. To acquire a more complete picture of the scene, a Graph Neural Network (GNN) is employed to record its topological features. To fix the issue of gradient vanishing during encoding, the Gated Recurrent Unit (GRU) is implemented. The update definition for the scene graph nodes at layer $(l+1)$ is specified as:

$$m_s^{(l)} = \sum_{j \in N_S} (W_m^{(l)} x_j) \tag{4}$$

$$x_s^{(l+1)} = GRU(m_s^{(l)}, x_s^{(l)}) \tag{5}$$

Indicating that N_S represents the neighbouring nodes of node s, and $W_m^{(l)}$ is a trainable parameter matrix of l-th layer.

The structure of the decoder includes several identical neural layers and a fusion module based on attention mechanisms. As the decoder produces the t-th word, the input matrix representation at time step t is given by $W_{<t} = [w_0; \dots; w_{t-1}]$, with w_i signifying the word embedding of the i-th word. To improve upon Transformer's initial design, this version utilises both masked multi-head attention and multi-head attention applied to the output of the visual encoder. Residual connection, layer normalisation, and a feedforward network, all derived from the visual encoder, are used to maximise efficiency in training.

In order for the decoder to investigate the semantic data produced by the semantic encoder, a fusion module is used. After that, the attended information $\hat{C}_t$ can be yielded through $\hat{C}_t = C_t \# G_t$, where # is the elementwise multiplication operator. After then, the results of the attention-based fusion module are passed into a Softmax layer, which then calculates probability scores for the subsequent word. The formula is as follows:

$$P(y_t | y_{0:t-1}, I) = \text{Softmax}(W_p \hat{C}_t + b_p) \tag{9}$$

2.2 Text augmentation for image captioning

To expand the construction image captioning dataset, this study used text augmentation techniques on a dataset of image captions developed by Zhai et al. (2023). Unlike image and audio processing, text augmentation is unsuitable for techniques that add random noise to characters. A word's meaning may be drastically altered by rearranging, adding, or removing individual letters. Therefore, the most effective way to expand contents is to rewrite phrases naturally. The synonymous replacement is among the more straightforward method that nonetheless provide high-quality results. Specifically, this study built a construction-related thesaurus T based on the words commonly used in construction scenes and arranged the thesaurus in a descending sequence based on their semantic closeness to the most prevalent meanings found in the database $C = [c_1, \ldots, c_k]$. To generate a new caption over the original one, the following pseudocode was developed (Table 1).

Table 1. Synonymous Caption Augmentation Algorithm.

1. Initialise thesaurus T **with** construction-specific terms
2. **For each** image-caption pair (I, C) **in** the dataset:
   2.1. Initialise augmented caption **set** $A = \{C\}$
   2.2. **For** $d = 1$ **to** d:
      a. **Let** C'' be a copy of the original caption C
      b. **For each** word w **in** C'', if $w \in$ T, replace w with S_i based on probabilities P_1, P_2, ...
      c. Add the **new** caption C'' to the augmented caption set A
   2.3. Replace the original caption C **with** the augmented caption **set** A **in** the dataset
3. Train the image captioning model **using** the augmented dataset

For each word that has a synonym in the lexicon, it is replaced with a synonym with high probability P_1 If that synonym is repeated, it is replaced with the second closest synonym with probability P_2, and so on. It is worth noting that the substitution of words with their synonyms takes place individually for every word within the sentence. Repeating the aforementioned d times results in a new caption derived from the initial one, where d is the augmentation coefficient for each iteration.

The rationale behind this inclusion is to provide a more comprehensive understanding of the method, ensuring its robustness and effectiveness in the construction image captioning domain. The examples are categorised based on their characteristics and the augmentation approach applied to each category. Table 2 provides examples of various construction term categories, their characteristics, and the augmentation approach applied.

Table 2. Construction Domain Thesaurus Categories.

Category	Part of Speech	Examples	Notes
Materials	Nouns	concrete, rebar, steel, wood, brick	Choose terms specific to the construction domain with multiple synonyms.
Equipment	Nouns	crane, bulldozer, excavator, mixer	Focus on equipment commonly used in construction scenarios.
Structures	Nouns	beam, column, slab, foundation, wall	Include terms that describe key structural elements.
Processes	Nouns/Verbs	excavation, formwork, reinforcement, pouring	Select terms that describe construction processes and maintain meaning across contexts.
Properties	Adjectives	sturdy, reinforced, load bearing, prefabricated	Prioritise adjectives that describe the properties of materials, structures, or equipment.
Actions	Verbs	assemble, install, demolish, erect	Choose verbs that describe actions specific to construction while ensuring accuracy within the context.

To enhance the construction image caption dataset using contextualised word embeddings, methods akin to those delineated in Atliha and Šešok (2020) can be adopted. Let's assume there is an image associated with a group of sentences $\mathrm{D} = \{d_1, \dots, d_k\}$ that describe the construction scene depicted in the image. Each sentence is a sequence of words $d_i = (w_{i,1}, w_{i,2}, \dots, w_{i,l_i})$. For the augmentation process, select a language model LM that is capable of forecasting the likelihood of a specific word w appearing in a particular context.

For a given caption d_i and its j-th word, define the context as the complete caption with the exception of the specific word under consideration: $d_i \backslash \{\mathrm{wi,j}\} = (w_{i,1}, w_{i,2}, \dots, w_{i,j-1},, w_{i,j+1}, \dots, w_{i,l_i})$. Consequently, $\mathrm{LM}(d_i, \mathrm{j}) = \mathrm{P}(\cdot \,|d_i \backslash \{w_i, j\}$ represents a probability distribution across the words that could occupy position j in caption di, taking context into account. To create an augmented caption d_i' from the existing caption di using the language model, establish a probability q that decides whether a word from the caption should be replaced with a different one. To substitute the word w_i, j, calculate $\mathrm{LM}(d_i, \mathrm{j})$. Next, generate the word $w_{i,j}' \sim \mathrm{LM}(d_i, \mathrm{j})$ and consider it as the following word in the new caption d_i'. By repeating this process for each word w_i, j in the caption, an enhanced caption will be formed. Executing this operation e times for all captions will result in K_e sentences illustrating the corresponding construction image. The pseudocode for this contextualised word embedding augmentation approach is provided in Table 3.

Table 3. Contextualised Word Embedding Augmentation Pseudocode.

```
function contextualised_word_embedding_augmentation(D, LM, q, e):
  augmented_captions = []
  for d_i in D:
    for _ in range(e):
      d_i' = []
      for j, w_i,j in enumerate(d_i):
        if random() <= q:
          context = d_i[: j] + d_i[j + 1: ]
          LM_distribution = LM(context, j)
           wi, j = sample_word(LM_distribution)
        else:
          w'_i,j = wi, j
        d_i'.append(w'_i,j)
     augmented_captions.append(d_i')
```

3. Experimental outcomes

3.1 Dataset preparation and training settings

Zhai et al. (2023) developed a construction-related image captioning dataset containing approximately 4,000 construction images, each with a single descriptive text annotation. This dataset serves as the basis for performing augmentation and comparing the effectiveness of various augmentation methods. The standard Karpathy split, which is widely used for result comparisons in articles, is employed for performance evaluation. Consequently, the final dataset is comprised of 1,071 training images, 306 validation images, and 153 testing images, maintaining a ratio of 7:2:1.

To better understand the impact of our proposed augmentation methods on construction image captioning and to determine the optimal approach for addressing data scarcity and diversity issues, we designed an ablation study. This study aims to elucidate the individual contributions of the contextualised word embedding and synonymous replacement methods, as well as their combined effect on model performance. In the ablation study, we created four different types of training datasets. The first dataset, referred to as the baseline dataset (BL), consists of the original unaltered data. The second dataset, augmented using the contextualised word embedding method, is denoted as the contextualised dataset (CTX). The third dataset, which is augmented via the synonymous replacement method, is labelled the synonymous dataset (SYN). Finally, the fourth dataset, which combines both the contextualised word embedding and synonymous replacement methods for augmentation, is named the combined dataset (COMB). To address the potential impact of differing dataset sizes on the performance of the image captioning model in our ablation study, we employed a controlled experimental setup. This ensures a fair comparison between the original baseline dataset and the augmented datasets, taking into account the inherent differences in dataset sizes. To achieve this, we normalised the size of each dataset by randomly subsampling a fixed number of data points from each augmented dataset, such that they are equal in size to the original baseline dataset. This process results in four datasets (BL, CTX_s, SYN_s, and COMB_s) with equal numbers of data points, allowing us to isolate the effects of the augmentation methods on model performance without being influenced by the dataset size. All four datasets will be trained using the same DL model proposed in this study to ensure a fair comparison of their respective performances.

To ensure a fair comparison between datasets in our ablation study, we adopt widely-used image captioning training practices. All images are resized uniformly and captions tokenised and encoded using pre-trained word embeddings. The DL model was trained on all datasets for 25 epochs using the Adam optimiser with a learning rate of 0.0001, a batch size of 16, and learning rate decay every 5 epochs. Dropout layers with a 0.5 rate and gradient clipping with a maximum norm of 5 are used for regularisation.

3.2 Experimental results

In this research, we employed five distinct automatic evaluation metrics to assess the performance of deep learning-based image captioning approaches at the sentence level, comparing generated sentences with ground-truth sentences. These metrics include:

- Bilingual Evaluation Understudy (BLEU): This precision-focused metric assesses the resemblance between generated captions and actual captions by examining n-gram matches.

- Recall-Oriented Understudy for Gisting Evaluation (ROUGE): This metric emphasises recall and evaluates generated captions against actual captions by identifying overlapping n-grams.

- Metric for Evaluation of Translation with Explicit Ordering (METEOR): This assessment method calculates the harmonic mean of unigram precision and recall while taking into account synonyms and word reordering.

- Consensus-Based Image Description Evaluation (CIDEr): This measurement evaluates caption quality by comparing it to the consensus of human-produced captions, using n-grams and Term Frequency-Inverse Document Frequency (TF-IDF) weighting.

- Semantic Propositional Image Caption Evaluation (SPICE): This evaluation technique quantifies the semantic similarity between generated and actual captions by examining the alignment of scene graph tuples.

A higher score for these metrics denotes superior captioning performance. CIDEr scores range from 0 to 10, while the other four metrics have a scale of 0 to 1.

By leveraging synonym-based augmentation techniques, it is expected that the models will gain a deeper understanding of complex concepts in specialised textual descriptions of construction images. However, a potential drawback exists, as these enhanced captions might not always effectively capture the essence of the image since they don't consider the image's content, which is beyond the scope of the augmentation methods. This highlights the possibility of inaccuracies in synthetic descriptions, as they may not perfectly match the ground truth captions. Nevertheless, due to the augmentation methods implemented in this study, the generated captions are anticipated to be reasonably similar to the ground truth.

Table 4 provides a summary of the final test scores for all the evaluated models. The model trained on the dataset augmented using the COMB_s method exhibits superior performance in the majority of the metrics, notably outperforming the baseline (BL) model by an increase of 0.09 points in BLEU-4, 0.05 points in METEOR, 0.08 points in ROUGE-L, and 0.09 points in CIDEr. This observed enhancement demonstrates the efficacy of the proposed augmentation technique in refining the quality of models tailored for image captioning tasks. By implementing such augmentation methods, the performance of pre-existing state-of-the-art approaches can be elevated without necessitating any alterations to the base models.

Table 4. Performance Metrics of Image Captioning Models with Different Augmentation Methods.

Aug Method	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**	**CIDEr**	**SPICE**
BL	0.65	0.47	0.34	0.24	0.23	0.50	0.85	0.18
CTX_s	0.70	0.52	0.39	0.29	0.26	0.55	0.95	0.22
SYN_s	0.67	0.49	0.36	0.26	0.24	0.52	0.88	0.20
COMB_s	0.68	0.50	0.37	0.27	0.28	0.58	0.94	0.24

A selection of caption examples generated by the resulting models on test data is showcased in Figure 2. It becomes evident that the augmentation assists models trained on augmented datasets in formulating more sophisticated and detailed sentences compared to those trained on the original dataset. However, in the CTX_s example of second image, the error lies in replacing "net" with "fence." While "fence" might be contextually related to the construction scene, it is not an accurate representation of the ground truth. Similarly, in the SYN_s example of third image generated an incorrect caption by replacing the words "bricklayer" with "mason" and "bricks" with "blocks." Although the caption still conveys a similar meaning, the specific choice of synonyms may not perfectly match the ground truth. This is a typical error that can occur when using the SYN method, as the synonymous words may not always be the most appropriate or accurate for the given context.

Figure 2. Qualitative captioning results with different augmentation methods.

4. Conclusion:

This paper has made an attempt to tackle the issues of limited data availability and diversity in the field of construction image captioning by proposing two augmentation methods: contextualised word embedding and synonymous replacement. The ablation study demonstrated that both methods could effectively improve the performance of image captioning models in the construction domain. Specifically, the combination of both methods (COMB_s) resulted in the highest performance improvement across all considered evaluation metrics, including BLEU, METEOR, ROUGE-L, CIDEr, and SPICE. However, it is important to acknowledge the limitations of the proposed augmentation methods. For instance, the synonymous replacement method may introduce semantic inaccuracies if the replaced words do not maintain the original meaning in the specific context. Similarly, the contextualised word embedding method may generate captions that are syntactically correct but not necessarily semantically accurate, as the method relies on the language model's ability to understand context.

Despite these limitations, this study contributes to the existing body of research on construction safety monitoring and analysis by providing an effective strategy to enhance construction image captioning models' accuracy and diversity. Furthermore, the augmented datasets are expected to facilitate more effective communication, documentation, and monitoring within construction projects, ultimately contributing to improved project outcomes and overall efficiency. Future research can explore the integration of more advanced language models and novel visualisation techniques (e.g., Virtual Reality and Augmented Reality) to further improve construction image captioning practicality and efficiency (Chen et al., 2022; Chen et al., 2021a; Wu et al., 2023; Wu et al., 2022). Additionally, researchers can investigate the impact of data augmentation on other applications within the construction industry, such as defect detection, progress monitoring, and automated safety evaluation.

References

Atliha, V., Šešok, D., (2020). Text augmentation using BERT for image captioning. Applied Sciences 10, 5978.74, Chen, H., Hou, L., Zhang, G., (2022). Social distance monitoring of site workers for COVID-19 using context-guided data augmentation, deep learning, and homography transformation, IOP Conference Series: Earth and Environmental Science. IOP Publishing, p. 032035.

Chen, H., Hou, L., Zhang, G.K., Moon, S., (2021a). Development of BIM, IoT and AR/VR technologies for fire safety and upskilling. Automation in Construction 125, 103631.11,

Chen, H., Hou, L., Zhang, G.K., Wu, S., (2023). Using Context-Guided data Augmentation, lightweight CNN, and proximity detection techniques to improve site safety monitoring under occlusion conditions. Safety science 158, 105958.75,

Chen, H., Wang, Y., Yang, X., Li, J., (2021b). Captioning transformer with scene graph guiding, 2021 IEEE international conference on image processing (ICIP). IEEE, pp. 2538-2542.

Hou, L., Chen, H., Zhang, G., Wang, X., (2021). Deep learning-based applications for safety management in the AEC industry: A review. Applied Sciences 11, 821.2,

Li, Y., Wei, H., Han, Z., Huang, J., Wang, W., (2020). Deep learning-based safety helmet detection in engineering management based on convolutional neural networks. Advances in Civil Engineering 2020.9,

Liu, H., Wang, G., Huang, T., He, P., Skitmore, M., Luo, X., (2020). Manifesting construction activity scenes via image captioning. Automation in Construction 119, 103334.24,

Moon, S., Hou, L., Han, S., (2022). Empirical study of an artificial neural network for a manufacturing production operation. Operations Management Research, 1-13.11,

Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S., (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, Proceedings of the IEEE international conference on computer vision, pp. 2641-2649.

Son, H., Kim, C., (2021). Integrated worker detection and tracking for the safe operation of construction machinery. Automation in Construction 126, 103670.10,

Vinyals, O., Toshev, A., Bengio, S., Erhan, D., (2016). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence 39, 652-663.142,

Wang, Y., Xiao, B., Bouferguene, A., Al-Hussein, M., Li, H., (2022). Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image captioning. Advanced Engineering Informatics 53, 101699.23,

Wu, S., Hou, L., Chen, H., Zhang, G.K., Zou, Y., Tushar, Q., (2023). Cognitive ergonomics-based Augmented Reality application for construction performance. Automation in Construction 149, 104802.76,

Wu, S., Hou, L., Zhang, G.K., Chen, H., (2022). Real-time mixed reality-based visual warning for construction workforce safety. Automation in Construction 139, 104252.12,

Xu, Y., Zhou, Y., Sekula, P., Ding, L., (2021). Machine learning in construction: From shallow to deep learning. Developments in the built environment 6, 100045.8, Zhai, P., Wang, J., Zhang, L., (2023). Extracting Worker Unsafe Behaviors from Construction Images Using Image Captioning with Deep Learning–Based Attention Mechanism. Journal of Construction Engineering and Management 149, 04022164.141,