Augmenting and Mixing Transformers with Synthetic Data for Image Captioning

Davide Caffagni[a], Marcella Cornia[b], Lorenzo Baraldi[a], Rita Cucchiara[a]

[a]*Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, Italy*
[b]*Department of Education and Humanities, University of Modena and Reggio Emilia, Reggio Emilia, Italy*

Abstract

Image captioning has attracted significant attention within the Computer Vision and Multimedia research domains, resulting in the development of effective methods for generating natural language descriptions of images. Concurrently, the rise of generative models has facilitated the production of highly realistic and high-quality images, particularly through recent advancements in latent diffusion models. In this paper, we propose to leverage the recent advances in Generative AI and create additional training data that can be effectively used to boost the performance of an image captioning model. Specifically, we combine real images with their synthetic counterparts generated by Stable Diffusion using a Mixup data augmentation technique to create novel training examples. Extensive experiments on the COCO dataset demonstrate the effectiveness of our solution in comparison to different baselines and state-of-the-art methods and validate the benefits of using synthetic data to augment the training stage of an image captioning model and improve the quality of the generated captions. Source code and trained models are publicly available at: `https://github.com/aimagelab/synthcap_pp`.

Keywords: Image Captioning, Synthetic Data, Vision-and-Language

1. Introduction

Image captioning presents complex challenges, as it involves describing images in natural language, bridging the realms of Computer Vision, Multimedia, and Natural Language Processing. Deep learning-based architectures for captioning have emerged as the dominant approach, setting a benchmark

in the field and becoming the go-to solution for new model designs. Despite significant progress, further improving the performance of these models has become increasingly difficult, primarily due to the limited availability of large, high-quality datasets containing sufficient image-caption pairs. To mitigate this challenge, existing methods typically rely on training captioning models [1, 2] using large-scale datasets sourced from the web [3, 4]. While web-sourced datasets provide a vast and diverse range of visual and textual knowledge, they introduce concerns related to content quality, ethical considerations, and the risk of misalignment between visual and textual information [5, 6, 7], which may hinder the performance and fairness of captioning systems.

A promising alternative for scaling modern neural networks while addressing the limitations of web-crawled data is the use of synthetic data. Synthetic data can be generated on demand in virtually unlimited quantities, eliminating annotation costs and offering greater flexibility in data creation. Additionally, it provides better control over content biases and ethical considerations compared to web-sourced datasets. These benefits position synthetic data as an effective solution for augmenting existing datasets or even substituting real data under specific constraints. While synthetic data has demonstrated significant potential across various Computer Vision tasks, such as object detection, segmentation, and tracking [8, 9, 10], its application in image captioning remains relatively underexplored.

Building on these premises, in this work we propose leveraging recent advancements in Generative AI [11, 12, 13, 14, 15] to enhance image captioning architectures. Specifically, we utilize the Stable Diffusion model [13], a well-established generative framework, to produce high-quality synthetic images aligned with human-annotated textual descriptions. These synthetic samples are then used to augment the widely-adopted COCO dataset [16], a standard benchmark for image captioning. To fully exploit the potential of synthetic data, we introduce a novel architecture, named SynthCap++, which extends the standard encoder-decoder Transformer framework and incorporates a specialized training strategy designed to integrate synthetic data effectively. In particular, we develop a Mixup-based [17] data augmentation technique that blends real and synthetic images at the pixel level, generating new visual samples that improve the generalization capability of the captioning model. We argue that the proposed Mixup synthetic augmentation is better suited for image captioning than traditional data augmentation such as rotation, flipping, cropping, jittering, and masking. Indeed, although these

techniques help in taming overfitting, they all perform a degree of perturbation on the source image. However, in image captioning, we must ensure that the augmented image preserves the coherence with its caption, which is hard to guarantee as augmentation is typically stochastic. Conversely, the proposed Mixup augmentation does not corrupt but rather substitutes the visual content of a real image with synthetic content which should be semantically coherent, under the hypothesis of a good image generator. Extensive experiments on the COCO dataset demonstrate the effectiveness of our approach, that consistently outperforms a baseline Transformer model without synthetic data augmentation and other competitive baselines specifically designed to validate our architectural and training choices. Moreover, the proposed model achieves competitive results even when compared with state-of-the-art models on both the standard COCO test set and the official online evaluation server, underscoring the potential of synthetic data in advancing image captioning performance.

Contributions. To sum up, the contributions of this paper are as follows:

- Thanks to the recent advances in generative architectures, we explore the benefits of using synthetically generated images to improve the training process of an image captioning model.
- To this aim, we design a new Mixup data augmentation strategy that effectively combines the pixels of real and synthetic images thus augmenting the data usually available to train an image captioning model and overall improving the generalization capabilities of the model.
- Experiments conducted on the COCO dataset demonstrate the effectiveness of our proposal, which achieves competitive performance also compared to state-of-the-art captioning models with more complex architectures and different carefully designed baselines.

This work is an improved and extended version of our conference paper [18]. With respect to this previous work, we introduce a new data augmentation strategy based on Mixup that differs from our previous approach in which we only exploit synthetic data as a replacement of real images with a certain probability at each iteration of the training process. We validate the usefulness of mixing real and synthetic images through extensive experiments and constantly demonstrate that SynthCap++ significantly improves the performance of the previous version.

2. Related Work

2.1. Image Captioning

Image captioning has come a long way since the early days of using predefined templates filled with relevant objects identified within images [19, 20]. The advent of deep learning has driven significant progress in the task, leading to the development of increasingly sophisticated models capable of producing more accurate and contextually rich captions. Early deep learning-based captioning frameworks predominantly followed a straightforward encoder-decoder paradigm [21, 22], wherein convolutional neural networks (CNNs) were employed to extract high-level features from input images, while recurrent neural networks handled the sequential generation of textual descriptions. Building on these initial efforts, subsequent techniques have progressively improved both the image encoding and language generation components of captioning models. One of the key advances in this direction has been the use of attention mechanisms [23] which allow the model to focus on specific parts of the image when generating a caption, either by computing attention over a grid of visual features [23] or utilizing image regions extracted from an object detector [24]. This has led to significant improvements in the accuracy and fluency of captions. Further improvements have been obtained with the use of semantic and spatial relationships encoded via graph neural networks [25, 26], which can provide additional information about the objects present on the scene and their interactions.

More recently, captioning literature has shifted towards the use of Transformer-based architectures [27, 28, 29, 30] initially designed for language understanding and machine translation tasks. These models have been used in captioning architecture with a dual role: first, in the visual encoding stage, where they refine features extracted from a CNN [31], an object detector [32, 33], or a Vision Transformer (ViT) [34], and second, as language models to generate descriptive captions [35, 36]. An alternative solution is to incorporate self-attention mechanisms to fuse visual features from multiple sources [37, 38, 39], such as a CNN and an object detector, even fine-tuning the visual backbones to increase the final performance [39]. Currently, a popular strategy involves utilizing visual features from CLIP-based [30] cross-modal architectures [40, 41, 42, 43], which has shown promising results in this domain. Moreover, these multimodal architectures offer the possibility of enhancing generated captions through retrieval components always based on CLIP features, as demonstrated in recent works [44, 45].

A different line of research capitalizes on the massive knowledge of Large Language Models and uses them to generate or refine image captions [46, 47]. Ultimately, Multimodal Large Language Models act as a multitask vision-language interface that can naturally be applied to image captioning [48, 49, 50]. In this work, we depart from using large-scale neural networks and target image captioning from a data perspective.

Considering instead the training phase, the typical optimization strategy employed in image captioning involves using time-wise cross-entropy loss [21, 23, 51]. Then, a reinforcement learning fine-tuning stage is almost always used to optimize the captioning model according to a specific non-differential metric (such as the CIDEr score [52] or other more recent metrics [53, 54, 55]), thus improving the final performance [22, 56, 57]. More recently, large-scale vision-and-language pre-training has been used to further enhance the final results, with the introduction of captioning models trained on million or even billion of image-text pairs collected from the web [58, 1, 2]. While these models have achieved outstanding performance, in this work we focus on training a Transformer-based captioning model on the COCO dataset alone by augmenting real images with synthetically generated data.

2.2. Data Augmentation

In the last years, data augmentation has gained attention as an effective way of virtually expanding the training dataset and reducing the overfitting typical of deep learning models. Usually, this is achieved by applying a perturbation on either the training samples, their labels, or both, increasing the diversity of training data and encouraging the model to generalize on the perturbed samples. While standard augmentation strategies typically transform input images via cropping, flipping, rotations, and other standard transformations, more sophisticated techniques focused on masking some areas of the images, randomly selected both in terms of size and position [59, 60]. Other solutions [17, 61, 62], instead, proposed to create new training samples by mixing the RGB values of two random training images according to a mixing factor sampled from a beta distribution, and also linearly combine the corresponding target labels using the same mixing factor. Following this strategy, known as Mixup [17], many other alternatives following the same principle have been introduced, either replacing a region of an image with another one taken from a different image [63] or mixing patches of two images in ViT-based architectures [64]. While these techniques have proven to be effective in many different tasks such as image classification [17], object detection [65],

and semantic segmentation [66], their role in image captioning is still unexplored. In this work, we draw inspiration from Mixup to effectively augment the training of a captioning model even though we restrict its application to the sole input, keeping the labels (*i.e.* the ground-truth captions) unaffected.

2.3. Synthetic Data in Vision-and-Language

Despite growing interest, research on leveraging synthetic data for image captioning remains relatively underexplored. Hossain *et al.* [67] pioneered the integration of synthetic images into captioning models by employing Generative Adversarial Networks (GANs) to synthesize novel visual inputs. However, the low fidelity and realism of images produced by early text-to-image GAN models limited their utility as an effective auxiliary training resource. More recent advancements by Xiao *et al.* [68] utilized a latent diffusion model [13] to augment the training set with generated images, accompanied by paraphrased textual descriptions. While this approach demonstrated improved realism, performance gains were notable primarily in low-data regimes or under unpaired image captioning scenarios. Li *et al.* [69] proposed fine-tuning a large-scale vision-language model by substituting difficult samples with synthetic counterparts, whereas Ma *et al.* [70] took a more radical approach by training exclusively on synthetic images. Their method employed a large language model to generate concise, context-relevant prompts for Stable Diffusion [13]. Building on these developments, our work also adopts a latent diffusion model for image synthesis. However, unlike prior methods, we rely solely on existing captions from the COCO dataset, requiring no additional textual augmentations and demonstrating the effectiveness of augmenting image captioning architectures with synthetic visual data. A different line of work [71, 72] studies how to generate high-quality synthetic captions to boost image captioning. Although we focus on synthetic images, these two strategies may have synergistic effects and could potentially be applied together.

3. Proposed Method

In this section, we introduce SynthCap++, a Transformer-based image captioning model that is augmented with synthetic images during training via a new data augmentation strategy based on Mixup [17]. Fig. 1 shows an overview of our model.

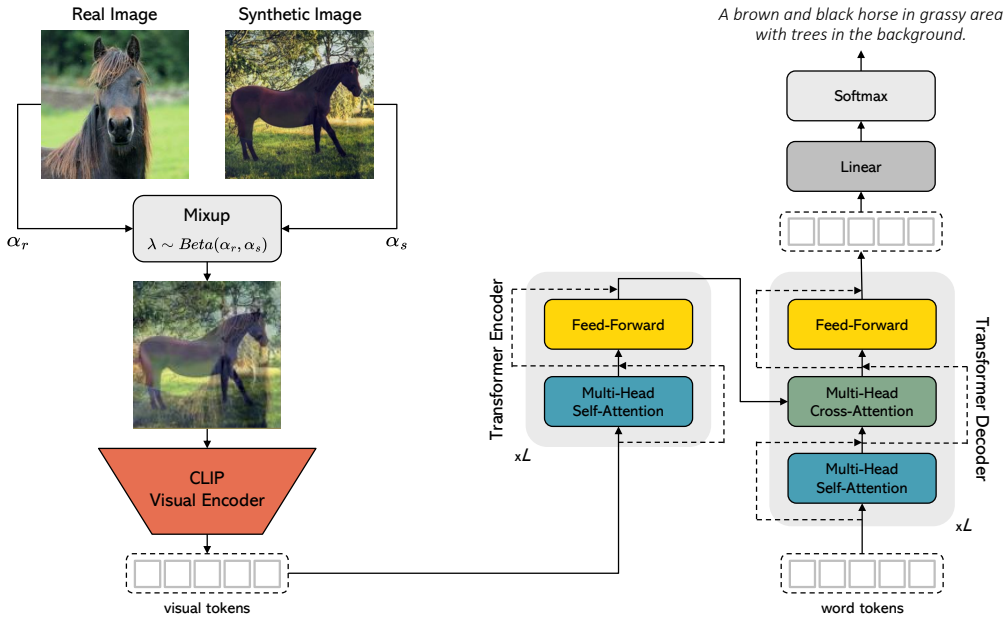

Figure 1: Overview of the proposed method. We first mix a real image along with one of its synthetic counterparts, according to a mixing factor drawn from a beta distribution with parameters α_r and α_s. Then, the CLIP visual encoder converts the mixed image into a sequence of visual tokens that are passed to the captioning model. Finally, a Transformer-based encoder-decoder architecture generates the output caption.

3.1. Model Architecture

Vision encoder. Our approach employs a fully-attentive Transformer architecture, designed to operate on visual features derived from a pre-trained visual encoder. Specifically, we leverage the image encoder from a CLIP-based model [30], keeping its parameters fixed across all experiments to ensure consistent feature extraction. As CLIP is trained to maximize alignment between images and corresponding textual descriptions, it offers robust visual representations that have demonstrated strong performance across various vision-language tasks, including image captioning [44, 42]. In this work, we employ the CLIP ViT-L/14 variant, which is based on the Vision Transformer (ViT) architecture [28]. This model divides input images into non-overlapping 14×14 patches, which are flattened and linearly transformed into a sequence of visual tokens, subsequently fed into the CLIP visual encoder.

Transformer. Our language model is based on a standard encoder-decoder Transformer architecture [27]. The encoder comprises multiple layers, each

consisting of a self-attention mechanism followed by a feed-forward network. The self-attention mechanism applies bi-directional attention to refine the visual tokens, while the feed-forward network processes each token independently using two fully connected layers separated by a GELU activation function [73]. Residual connections [74] are incorporated by adding the output of each block to its input, followed by layer normalization [75].

The decoder adopts a similar structure but incorporates an additional cross-attention mechanism between the self-attention and feed-forward blocks. This cross-attention module is critical for fusing information across visual and textual modalities, where the partial caption tokens generated up to time step t serve as queries, attending to the visual tokens (used as keys and values) from the encoder. To enforce autoregressive generation, a causal mask is applied to the decoder self-attention, ensuring that tokens only attend to past or current positions by masking out future positions – specifically, entries with indices $(i, j)_{\forall j>i}$ in the attention matrix are zeroed.

The decoder produces a token sequence $\widetilde{\mathbf{x}} = \{\widetilde{x}_t\}_{t=1,\ldots,N}$ of the same length as the input. The next token $\widetilde{x}_{t+1}$ is sampled from a probability distribution over the vocabulary, obtained by applying a linear transformation followed by a softmax activation to $\widetilde{x}_t$. During inference, the decoder operates in an autoregressive manner, where the output token at each time step t is fed back as input for the subsequent step $t+1$, enabling sequential caption generation.

3.2. Synthetic Data Mixing

Our objective is to investigate the potential of synthetic images as a valuable source for training captioning algorithms. To achieve this, we employ the Stable Diffusion model [13] to generate synthetic images that expand the training set of the COCO dataset [51]. The original COCO dataset contains over half a million image-caption pairs (I^r, c_k), where each image I^r has five different reference descriptions c_k with $k = 1, 2, 3, 4, 5$. By conditioning the Stable Diffusion model on each caption c_k, we generate an additional dataset comprising synthetic image-caption pairs (I^s_k, c_k). Our experimental analysis reveals that the generated synthetic images maintain a high degree of semantic alignment with their corresponding captions, underscoring the potential of synthetic data as an effective augmentation strategy for image captioning. Nevertheless, we observe that training a model exclusively on synthetic images and their paired captions results in suboptimal performance. Consequently, we argue that real and synthetic images play complementary roles in image captioning, and leveraging both can enhance model performance.

Following this premise, we propose to combine real and synthetic images in the same forward pass, instead of simply selecting a real or a synthetic sample in a mutually exclusive manner (as done in our previous work [18]). To this end, we design a new data augmentation strategy based on Mixup [17] to combine real-synthetic image pairs at the pixel level and obtain new training samples. Formally, given a triplet (I^r, I^s_k, c_k), we create a new image-caption pair (I, c_k), whose visual part is created according to the Mixup data augmentation strategy that linearly interpolates the pixels of the two images. Formally, the new image I is created as follows:

$$I = \lambda I^r + (1 - \lambda) I^s_k, \tag{1}$$

$$\lambda \sim Beta(\alpha_r, \alpha_s), \tag{2}$$

where α_r and α_s are two parameters that regulates the distribution $Beta(\alpha_r, \alpha_s)$ from which λ is sampled at each iteration. Intuitively, as λ increases, I^r takes priority over I^s_k. The beta distribution parameters can be considered as filters to control the average amount of information drawn from the real and the synthetic domains. In the experiments, we show how the performance changes as the α parameters vary.

It is important to notice that the original Mixup data augmentation strategy corrupts the visual input given to the model. In the context of image captioning, we need that such corruption still preserves the coherence of the input image with respect to the ground-truth caption. Thus, using a synthetic image generated from the same caption associated with its real counterpart guarantees preserving the same visual content when combining the two images. Moreover, Mixup is usually applied on both the input and the target labels, using the same λ parameter for each sample. In our setup, however, we have a single ground-truth caption that describes a specific real-synthetic pair. Therefore, we decide to leave the labels untouched and apply Mixup only on the input.

3.3. Training Procedure

To train our model, we adopt the conventional two-phase training procedure commonly employed in image captioning [24, 36, 32]. The process begins with a pre-training phase utilizing cross-entropy loss, followed by fine-tuning through self-critical sequence training (SCST) [22]. SCST leverages reinforcement learning, where the CIDEr metric [52] serves as the reward signal to optimize the captioning model.

During SCST optimization, we follow standard practice as demonstrated in prior works [36], where the baseline reward is computed as the average score across all sequences sampled using beam search within the same beam. In this framework, we apply Mixup between a real-synthetic image pair by randomly selecting one of the five synthetic images associated with the real image. Formally,

$$I_k^s \sim \{I_1^s, I_2^s, I_3^s, I_4^s, I_5^s\} \ . \tag{3}$$

It is important to note that, although for each k, the synthetic image I_k^s is generated from a single description c_k, the CIDEr metric still evaluates the consensus of the captions produced by our model against all five reference captions c_k with $k = 1, 2, 3, 4, 5$.

4. Experimental Evaluation

In this section, we present analyses and experiments conducted to validate the effectiveness of the proposed SynthCap++ model and the comparison with state-of-the-art methods. First, we describe the dataset, evaluation metrics, and the implementation and training details used in our experiments.

4.1. Experimental Setting

Dataset. We validate our proposal on the Microsoft COCO dataset [16], complying with the standard Karpathy splits [51]. Specifically, the dataset contains more than 120k images annotated with five different captions each, where 5,000 images are used for validation, 5,000 for testing, and the rest for training. In our experiments, we test our solution on both the standard test set and on the COCO online test server composed of more than 40k images for which ground-truth captions are not publicly available. To augment the dataset with synthetic images, we generate images with Stable Diffusion [13] using the 1.4 version provided by the Huggingface library[1]. In particular, each caption of the original COCO dataset is used to generate the corresponding synthetic image using the generic prompt "*An image of*" at the beginning of the caption. During image generation, the safety checker module was employed to reduce the probability of explicit images and the invisible watermarking of the outputs was disabled to prevent identification of synthetic images as machine-generated.

[1] `https://huggingface.co/CompVis/stable-diffusion-v1-4`

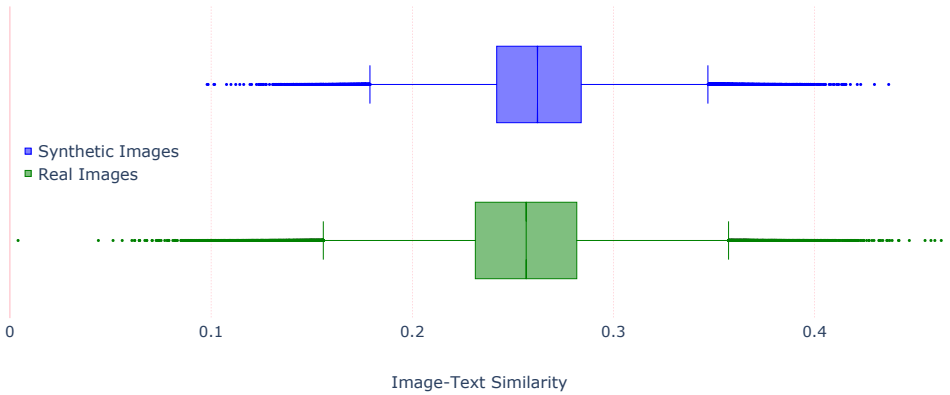

Figure 2: Image-text similarity score computed by CLIP with real and synthetic images.

Evaluation metrics. Evaluation is done according to the usual metrics for image captioning, namely BLEU (B-1 and B-4) [76], METEOR (M) [77], ROUGE (R) [78], CIDEr (C) [52], and SPICE (S) [79].

Implementation details. Before being fed into the CLIP visual encoder, each image undergoes a pre-processing pipeline. Initially, the image is resized to ensure that the longer side does not exceed 224 pixels, while maintaining the original aspect ratio. Subsequently, the image is center-cropped and normalized on a per-channel basis. This results in an input tensor with dimensions $3 \times 224 \times 224$. From this input, the ViT-based CLIP encoder extracts a grid of 256×1024 features, which correspond to the visual tokens. Our Transformer-based image captioning network consists of an encoder and a decoder, each comprising $L = 3$ layers with a hidden size d equal to 512. To match this dimensionality, a linear projection is applied to the CLIP visual features. The attention layers employ multi-head attention with 8 heads, accompanied by a dropout rate of 0.1. For tokenizing words, we use the same byte-pair encoding (BPE) tokenizer [80] as the CLIP textual encoder.

Training details. We pre-train our model using cross-entropy loss for 30 epochs, selecting the checkpoint with the highest CIDEr score. Subsequently, we fine-tune the model via reinforcement learning within the SCST framework [22]. During the cross-entropy phase, we adopt the setup from [44], using a batch size of 32 and a learning rate schedule based on [27], with a warmup of $20,000$ iterations. In the SCST phase, we set the batch size to 16, use a fixed learning rate of 10^{-6}, and apply beam search decoding with a beam size of 5. For both phases, we use the Adam optimizer [81].

All experiments are conducted with mixed precision [82] and ZeRO memory offloading [83], utilizing the Huggingface Transformers library [84]. For computational efficiency, we always keep the visual encoder frozen, and only optimize the weights of the Transformer encoder-decoder modules. We believe that our method may further benefit from unfreezing the visual encoder, but we leave this experiments to future developments.

Mixup details. As described in Section 3.2, we mix only real and synthetic images, keeping the labels (*i.e.* the reference captions) untouched. The coefficients α_r and α_s are fixed and frozen at the beginning of each training run. For efficiency, we sample a single λ value for each batch from the beta distribution and apply the corresponding Mixup operation to all the samples contained in the batch. In our experiments, we notice that this does not affect the final performance.

4.2. Ablation Studies and Analyses

Overall validation of synthetic images. As an initial analysis, we evaluate the alignment between synthetic images and their corresponding captions to assess the potential of synthetic images as a valuable source for training image captioning models. To do that, we compute the cosine similarity between each image-text pair using the same ViT-based CLIP variant used to encode images during training (*i.e.* ViT-L/14). In particular, the similarity score is computed by feeding image-text pairs to the CLIP visual and textual encoders and computing the cosine similarity between the two embedded representations. As shown in recent literature [53], this can be a valuable metric to measure the correspondence of an image with a given textual sentence. The computed scores are reported in Fig. 2, where we compare image-text similarities of real and synthetic images with the corresponding captions, respectively. On average, synthetic images have a higher affinity to their captions than the original ones, thus suggesting that they can be a good training source for an image captioning model.

Mixup coefficients. We then analyze in Table 1 the best configuration of Mixup coefficients. We compare the results with a baseline model that follows the exact same Transformer-based architecture used in our model but that is trained on real images only. We start with the Mixup coefficients proposed in the original implementation (*i.e.* $\alpha_r = 0.4$ and $\alpha_s = 0.4$). Although it is interesting to experience some minor improvements against the baseline (*i.e.* 127.2 CIDEr points vs. 126.5 of the baseline), this setup is probably too

Synth. Data	# Synth. Ims	Augmentation	α_r	α_s	B-1	B-4	M	R	C	S
✗	-	-	-	-	77.5	37.2	30.0	58.6	126.5	23.3
✓ (only)	1	-	-	-	72.7	29.2	25.5	53.1	100.2	19.0
✗	-	Standard	-	-	77.9	38.0	30.5	58.9	128.0	23.5
✓	1	CutMix	0.9	0.1	77.6	37.8	30.4	59.0	128.1	23.3
✓	1	TokenMix	0.9	0.1	77.3	37.3	30.4	58.9	127.6	23.4
✓	1	Mixup (img+text)	0.9	0.1	77.3	37.5	**30.7**	58.3	127.8	23.5
✓	1	Mixup	0.4	0.4	76.9	37.5	30.1	58.5	127.2	23.4
✓	1	Mixup	0.6	0.4	77.7	37.7	30.4	58.9	128.0	**23.6**
✓	1	Mixup	0.7	0.3	77.2	37.0	30.0	58.6	127.5	23.4
✓	1	Mixup	0.8	0.2	77.3	37.3	30.2	58.9	127.4	**23.6**
✓	1	Mixup	0.9	0.1	**78.0**	**38.1**	30.6	**59.3**	**129.0**	**23.6**
✓	1	Mixup	0.95	0.05	77.8	37.9	30.6	59.1	127.8	23.4
✓	2	Mixup	0.9	0.1	77.4	37.7	**30.7**	58.3	128.0	23.3
✓	5	Mixup	0.9	0.1	77.6	37.5	30.4	58.2	127.4	23.0

Table 1: Ablation study using different augmentation strategies and α values on the COCO Karpathy test set. Results are reported after cross-entropy pre-training. Higher is better for all metrics (↑).

severe. In fact, since the mean λ value sampled from the beta distribution is 0.5, the model rarely sees non-augmented real images, which however is important to preserve the generalization to the natural domain. Therefore, we increase α_r at the expense of α_s in subsequent experiments, thus prioritizing real images. As it can be seen, as α_r increases, the final results generally improve, with the highest CIDEr score obtained with α_r equal to 0.9 and α_s equal to 0.1. We can also notice that further increasing α_r and consequently decreasing α_s leads to worse performance. In contrast, training the captioning model solely on synthetic images results in a notable performance drop compared to the baseline. This degradation can be attributed to the reality gap between real and synthetic images, which hinders the ability of the model to generalize to real data when trained exclusively on synthetic samples.

Alternative data augmentation strategies. To demonstrate the effectiveness of our Mixup strategy, we compare it with other alternatives typically used for data augmentation. In particular, we employ two mixing-based strategies, namely CutMix [63] and TokenMix [64]. Rather than a convex combination between two images, these methods work with cut-and-paste operations. CutMix does such an action once, while TokenMix performs multiple moves, where each cut interests a patch that corresponds to a token

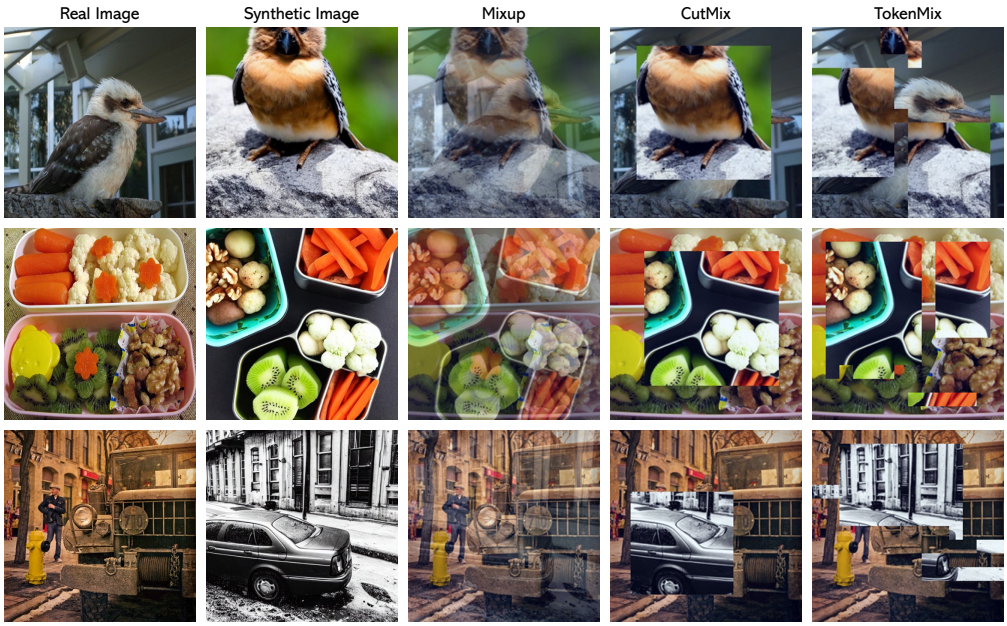

Figure 3: Qualitative results of the application of the proposed Mixup data augmentation strategy in comparison with the CutMix and TokenMix variants.

inside a ViT-based architecture. When implementing TokenMix, we remove the constraint on any minimum number of patches to be mixed. This means that, during training, the model may still see full real images, if the sampled number of mixing patches is 0. In CutMix, the coefficients α_r and α_s control the average size of the synthetic crop that is pasted on the real image. When $\alpha_r > \alpha_s$, the model on average sees images where the synthetic crop covers a smaller percentage of pixels compared to the real image, and vice versa. Similarly, in TokenMix, those coefficients modulate the average amount of real pixel patches that will be substituted by synthetic ones. The substitution is done at the pixel level, right before the final image gets embedded into visual tokens by the patch embedding layer of the vision transformer.

Results are reported in Table 1. While still improving over the baseline, they underperform compared to Mixup, which achieves 129.0 CIDEr points against 128.1 and 127.6 respectively for CutMix and TokenMix. This is partially in contrast with the results obtained by image classification literature, where CutMix and TokenMix perform better than a standard Mixup data augmentation strategy. However, our setting is not directly comparable with the one employed in image classification for two main reasons: (i) we do not

take into account the target labels, which remain unaltered, and (ii) we mix real images with synthetic ones, a strategy not yet explored in the literature. To provide a qualitative visualization of the compared mixing techniques, we report in Fig. 3 some qualitative examples of the application of these strategies on sample real-synthetic image pairs.

As an additional analysis, we also investigate a possible extension of Mixup accounting for text. Specifically, during cross-entropy training, we linearly interpolate pixels from I^r and I^s_k, as well as the text embeddings obtained from caption $c_{j,j\neq k}$ and c_k using the input matrix embeddings of the Transformer decoder. We also interpolate the target probabilities with the same coefficients. While improving over the baseline on BLEU-4, METEOR, CIDEr, and SPICE, employing Mixup on both images and captions underperforms against mixing images only.

To further validate the effectiveness of our Mixup strategy, we also compare against a model trained with standard data augmentation operations, without using synthetic data as an additional training source. In this configuration, we feed the model with the original image with a probability of 0.5, and we apply a random rotation of $\pm 30°$ or a horizontal flip, both with a probability of 0.25. Results are better than those achieved by the baseline without data augmentation but are still far from those obtained with the best Mixup-based model, thus further demonstrating the effectiveness of our augmentation strategy.

Varying the number of synthetic images. The last two rows of Table 1 show the results of our best Mixup configuration, when mixing a real image with multiple synthetic images. For these experiments, we modify Eq. 2 to account for $n > 1$ synthetic images, as follows:

$$I = \lambda I^r + \frac{1}{n}\sum_{i=1}^{n}(1-\lambda)I^s_i. \tag{4}$$

In this setting the ground-truth caption is randomly chosen among the five available. However, the benefits of this augmentation strategy are not as robust as those of the standard Mixup scheme with $n = 1$: METEOR and CIDEr are higher than the baseline, while other metrics are on par or slightly lower. Moreover, the experiment with $n = 2$ is superior than that with $n = 5$, further suggesting that the number of synthetic images to be mixed should be limited.

	Synth. Data	α_r	α_s	B-1	B-4	M	R	C	S
Transformer	✗	-	-	77.5	37.2	30.0	58.6	126.5	23.3
Transformer + Mixup (w/ similar ims, $k = 1$)	✗	0.6	0.4	77.0	37.5	30.6	58.2	127.3	23.1
Transformer + Mixup (w/ similar ims, $k = 1$)	✗	0.7	0.3	77.6	37.6	30.6	58.3	127.3	23.3
Transformer + Mixup (w/ similar ims, $k = 1$)	✗	0.8	0.2	77.9	**38.1**	**30.7**	58.7	127.6	23.1
Transformer + Mixup (w/ similar ims, $k = 1$)	✗	0.9	0.1	77.3	37.4	30.1	58.8	126.8	23.5
Transformer + Mixup (w/ similar ims, $k = 3$)	✗	0.9	0.1	76.9	36.9	30.0	58.4	125.5	23.2
Transformer + Mixup (w/ similar ims, $k = 5$)	✗	0.9	0.1	77.3	37.3	30.1	58.7	125.8	23.1
SynthCap [18]	✓	-	-	77.7	37.6	30.3	58.9	128.6	23.4
SynthCap++	✓	0.9	0.1	**78.0**	**38.1**	30.6	**59.3**	**129.0**	**23.6**

Table 2: Analysis on the effectiveness of using synthetic data during training. Results are reported after cross-entropy pre-training on the COCO Karpathy test set. Higher is better for all metrics (↑).

Effectiveness of synthetic data. To verify that the observed improvements can be directly attributable to the use of synthetic images for augmenting our training set, rather than being a mere consequence of Mixup, we modify the source of visual input used for augmentation. Since synthetic images are inherently similar to their original counterparts, it is more appropriate to compare them with real images that share visual similarity. Thus, in this case, given an image I^r from the COCO dataset, we mix it with $\tilde{I}^r_k$, that corresponds to a real image randomly selected among the top-k similar images with respect to I^r. To identify these similar images, we first extract feature vectors for each image using a pre-trained CLIP model. For a given query image, the k most similar images are retrieved, where k can be set to 1, 3, or 5, based on cosine similarity between the feature vectors as the similarity measure. In this case, Mixup is performed with the best configuration obtained in the previous experiments (*i.e.* $\alpha_r = 0.9$, $\alpha_s = 0.1$). As it can be noticed from the results shown in Table 2, using synthetic images to perform Mixup leads to the best performance. For completeness, we also report the results of the previous version of our model (*i.e.* SynthCap [18]) which were obtained by replacing real images with synthetic ones during training, without performing Mixup. When comparing the two versions of our model, we can see that SynthCap++ achieves the best results according to all evaluation metrics.

To complete the analysis, the first three rows of Table 2 show the results of varying the mixing coefficient for the case $k = 1$, as it is the best configuration when mixing real images with the same coefficients of SynthCap++. While some performance gains are observed in certain configurations, the

	Synth. Data	Backbone	B-1	B-4	M	R	C	S
Transformer	✗	DINOv2 ViT-L/14	76.7	36.7	30.1	57.6	124.3	22.9
SynthCap++	✓	DINOv2 ViT-L/14	**76.9**	**37.1**	**30.4**	**58.0**	**125.1**	**23.2**
Transformer	✗	CLIP ViT-B/32	**75.1**	34.3	29.0	56.0	114.8	21.4
SynthCap++	✓	CLIP ViT-B/32	74.3	**34.7**	**29.5**	**56.4**	**115.6**	**21.7**
Transformer	✗	OpenCLIP ViT-L/14	75.8	35.8	29.8	57.1	121.5	22.4
SynthCap++	✓	OpenCLIP ViT-L/14	**76.5**	**36.2**	**30.1**	**57.5**	**122.4**	**22.7**
Transformer	✗	CLIP ViT-L/14	77.5	37.2	30.0	58.6	126.5	23.3
SynthCap++	✓	CLIP ViT-L/14	**78.0**	**38.1**	**30.6**	**59.3**	**129.0**	**23.6**

Table 3: Comparison with different visual encoders. Higher is better for all metrics (↑).

use of synthetic images consistently yields superior performance across most evaluation metrics.

Effect of Mixup using diverse visual encoder. Table 3 outlines how SynthCap++ can generalize to different visual encoders other than CLIP ViT-L/14. All experiments are executed with and without our best synthetic augmentation scheme, that is Mixup with $\alpha_r = 0.9$ and $\alpha_s = 0.1$. In the first two rows, we employ DINOv2 ViT-L/14 [85], a Vision Transformer trained with self-supervised learning on images only. As it can be seen, SynthCap++ consistently improves over all the considered metrics, testifying its efficacy beyond visual encoders pretrained with natural language supervision. Concerning this latter family of encoders, we experiment with a smaller version of CLIP, featuring a ViT-B/32 model. Also at this scale, SynthCap++ leads to better results across all but the BLEU-1 metric. Finally (last and second rows), we include OpenCLIP ViT-L/14 [86], that closely follows the architecture and pre-training objective of CLIP, but which has been trained on LAION-2B [4]. While it falls short of the original CLIP, SynthCap++ records noticeable gains on all metrics with respect to the baseline.

Computational analysis. We report runtime and VRAM statistics averaged over 1,000 training/test samples. We run these experiments on a single NVIDIA 2080 Ti 11GB GPU, using CLIP ViT-L/14 as the visual encoder. We implement Mixup, CutMix, and TokenMix straight inside the `forward` method of the model, to benefit from CUDA acceleration and simplify runtime assessment. During the cross-entropy stage, single forward and forward-backward passes with batch size 32 take 155 ms and 194 ms respectively for the baseline Transformer. The impact of augmentation with Mixup schemes is minimal, as SynthCap++ records 161 ms (forward) and 194 ms (forward-

	Cross-Entropy Loss						CIDEr Optimization					
	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
Up-Down [24]	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
GCN-LSTM [25]	77.3	36.8	27.9	57.0	116.3	20.9	80.9	38.3	28.6	58.5	128.7	22.1
SGAE [26]	77.6	36.9	27.7	57.2	116.7	20.9	81.0	39.0	28.4	58.9	129.1	22.2
AoANet [32]	77.4	37.2	28.4	57.5	119.8	21.3	80.2	38.9	29.2	58.8	129.8	22.4
$\mathcal{M}^2$ Transformer [36]	-	-	-	-	-	-	80.8	39.1	29.2	58.6	131.2	22.6
X-Transformer [33]	77.3	37.0	28.7	57.5	120.0	21.8	80.9	39.7	29.5	59.1	132.8	23.4
DLCT [37]	-	-	-	-	-	-	81.4	39.8	29.5	59.1	133.8	23.0
RSTNet [31]	-	-	-	-	-	-	81.8	40.1	29.8	59.5	135.6	23.3
DIFNet [38]	-	-	-	-	-	-	81.7	40.0	29.7	59.4	136.2	23.2
CaMEL [42]	78.3	39.1	29.4	58.5	125.7	22.2	82.8	41.3	30.2	60.1	140.6	23.9
COS-Net [44]	79.2	39.2	29.7	58.9	127.4	22.7	82.7	42.0	30.6	60.6	141.1	24.6
GRIT† [39]	-	-	-	-	-	-	**84.2**	42.4	30.6	60.7	144.2	24.3
Transformer	77.5	37.2	30.0	58.6	126.5	23.3	82.9	42.2	30.7	60.9	141.9	24.6
SynthCap [18]	77.7	37.6	30.3	58.9	128.6	23.4	83.0	42.4	30.8	61.1	143.1	24.7
SynthCap++	**78.0**	**38.1**	**30.6**	**59.3**	**129.0**	**23.6**	83.5	**42.9**	**31.2**	**61.5**	**144.4**	**25.1**

Table 4: Comparison with the state of the art on the COCO Karpathy test set. The † marker indicates fine-tuning of the visual backbone. Higher is better for all metrics (↑).

backward). The effect of switching from Mixup to CutMix or TokenMix is neglectable. GPU VRAM is steady at 4.8 GB. Because Mixup schemes only play a role during training, at inference time, SynthCap++ shares the same computational costs as the Transformer baseline. A single caption generation takes 154 ms and up to 1.8 GB VRAM.

4.3. Comparison to the State of the Art

We then perform a comparison with state-of-the-art captioning models in Table 4. Specifically, we include methods based on recurrent neural networks with attention mechanisms over regions such as Up-Down [24], eventually enriched with spatial and semantic graphs (*i.e.* GCN-LSTM [25] and SGAE [26]) and with self-attention operators (*i.e.* AoANet [32]). The comparison also comprises more recent Transformer proposals like $\mathcal{M}^2$ Transformer [36], X-Transformer [33], DLCT [37], RSTNet [31], and DIFNet [38]. Finally, we include CaMEL [42] and COS-Net [44] that both employ CLIP-based features, where the latter also retrieves knowledge from an external base, and GRIT [39] that instead is based on a combination of features extracted from different backbones which are fine-tuned during training. In this case, the results are reported after both the first training stage with cross-entropy loss and the second one based on reinforcement learning with CIDEr optimization. We also add in the comparison a standard Transformer model

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Up-Down [24]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
SGAE [26]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
AoANet [32]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
$\mathcal{M}^2$ Transformer [36]	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
X-Transformer [33]	81.9	95.7	66.9	90.5	52.4	82.5	40.3	72.4	29.6	39.2	59.5	75.0	131.1	133.5
RSTNet [31]	82.1	96.4	67.0	91.3	52.2	83.0	40.0	73.1	29.6	39.1	59.5	74.6	131.9	134.0
DLCT [37]	82.4	96.6	67.4	91.7	52.8	83.8	40.6	74.0	29.8	39.6	59.8	75.3	133.3	135.4
COS-Net [44]	83.3	96.8	68.6	92.3	54.2	84.5	42.0	74.7	30.4	40.1	60.6	76.4	136.7	138.3
CaMEL [42]	83.2	97.3	68.3	92.7	53.6	84.8	41.2	74.9	30.2	39.7	60.2	75.6	137.5	140.0
GRIT† [39]	**84.1**	97.6	69.4	93.5	54.9	86.3	42.5	76.8	30.9	41.0	61.2	77.1	**141.3**	**143.8**
SynthCap [18]	83.7	97.6	69.2	93.5	54.9	86.3	42.8	77.1	30.9	41.3	61.4	77.7	140.1	142.6
SynthCap++	**84.1**	**97.7**	**69.6**	**93.7**	**55.3**	**86.6**	**43.2**	**77.4**	**31.1**	**41.4**	**61.6**	**77.9**	141.2	143.2

Table 5: Leaderboard of various methods on the online COCO test server. The † marker indicates fine-tuning of the visual backbone. Higher is better for all metrics (↑).

without any data augmentation and the previous version of our architecture that does not include Mixup. Both these models have the same configuration in terms of visual features and architecture used in SynthCap++ and are trained with the same hyperparameters.

As it can be seen, SynthCap++ overcomes all competitors in both the cross-entropy training phase and CIDEr-based optimization, achieving the best results on almost all metrics. In particular, our model reaches an improvement of 2.5 CIDEr points compared to the Transformer baseline in both training stages confirming the usefulness of augmenting the training data with synthetic images. When compared with its previous version, SynthCap++ exhibits better performance in both settings with an improvement of 1.2 CIDEr points after CIDEr-based optimization, demonstrating the benefits of using Mixup in combination with synthetic images as data augmentation strategy. It is worth noting that our model is also competitive with respect to GRIT [39], whose performance is boosted by fine-tuned visual backbones.

We also present in Table 5 the comparison on the official COCO test split, through the online test server. Results are reported in terms of the standard captioning metrics with respect to 5 reference captions (c5) and 40 reference captions (c40). Following previous works [36, 44], we compute the final predictions using an ensemble of four models, obtained by training each model with a different random seed. Also in this setting, SynthCap++ achieves the best results according to almost all metrics, except GRIT [39]

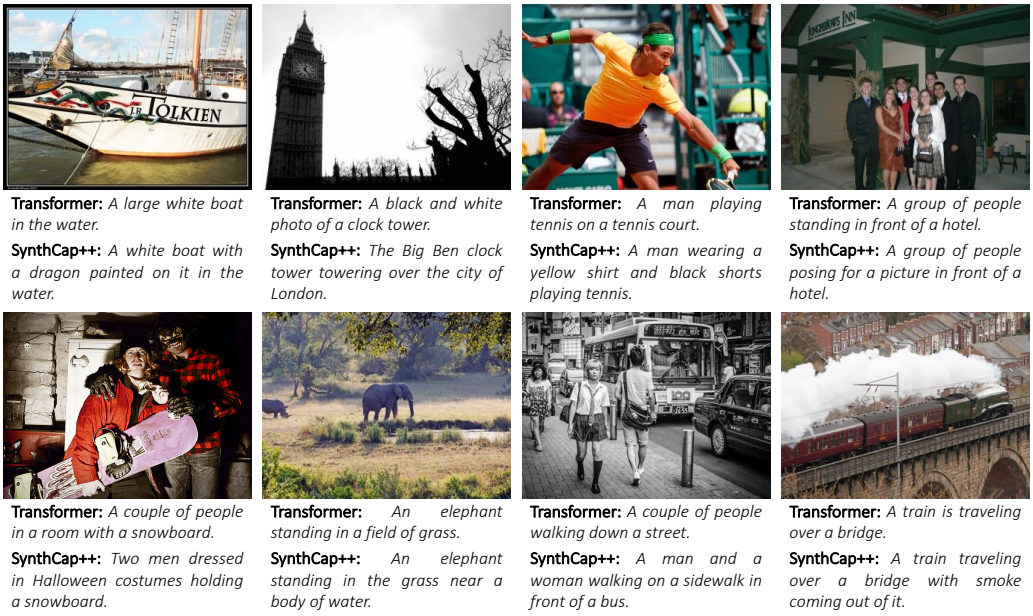

Figure 4: Sample captions generated by SynthCap++ describing images from the COCO dataset. For comparison, we also report captions generated by a standard Transformer model without synthetic data augmentation.

which achieves slightly better results in terms of CIDEr scores, and also overcomes the performance of the previous version (*i.e.* SynthCap [18]).

4.4. Qualitative Results

Finally, some qualitative results are shown in Fig. 4 comparing captions generated by our method against those generated by the standard Transformer model on sample images from the COCO dataset. As it can be seen, captions predicted by SynthCap++ are generally more detailed and better describe the visual content of input images, correctly identifying some specific concepts reported in the images like the "Big Ben" and the "city of London" in the second example of the first row and the "Halloween costumes" in the first example of the second row. Also, even if the captions predicted by the baseline are generally correct, SynthCap++ can describe finer-grained details such as the "dragon" painted on the boat shown in the first example of the first row and the "smoke" coming out of the train depicted in the last image of the second row.

Figure 5: Examples of erroneous or incomplete captions generated by SynthCap++.

4.5. Limitations and Failure Cases

While SynthCap++ provides an effective framework for state-of-the-art image captioning results, it may also have some limitations. First, the proposed solution requires the generation of synthetic images, that introduces additional computational cost and makes the model tightly bound to the quality of the image generator. Second, although our experiments (cf. Table 1) show that a mix of real and synthetic data is necessary for effective training, determining optimal Mixup coefficients remains non-trivial. In the absence of prior work using Mixup with synthetic data for image captioning, we empirically selected the coefficients yielding the best results on COCO. However, this process can be computationally expensive for larger datasets and may depend on the specific image generator used.

Finally, like all captioners, SynthCap++ is not free from mistakes, as depicted in Fig. 5. For instance, it may mislabel a *table* instead of a bin or wooden basket; confuse the location of subjects (*e.g.*, dogs on the passenger seat instead of the *driver*'s seat); or make counting errors, such as referencing *two* children when only one is present. Even when correctly identifying visible entities, as in the fourth image, it can overlook partially occluded ones, such as the *young girl* behind the front man. A possible explanation for such errors is that Mixup introduce visual artifacts while linearly interpolating two images, potentially misleading the model into hallucinating elements not present in the original images.

5. Conclusion

In this work, we proposed SynthCap++, a new architecture for image captioning that leverages synthetic visual data generated by Stable Diffusion

during training. In particular, our model relies on Mixup as a data augmentation strategy to combine pixels from real and synthetic distributions into new images that are used as training samples for the model along with the corresponding captions. Experiments, conducted on the COCO dataset in comparison with state-of-the-art methods and different baselines, demonstrate the effectiveness of our proposal and validate the benefits of using synthetic images as a valuable source of information during training, as well as the utility of mixing them with real images. We believe this could be a first step towards using synthetic data to improve the performance of an image captioning model, possibly leading to further research in this direction.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This work has been supported by the PNRR project "Italian Strengthening of Esfri RI Resilience (ITSERR)" (CUP B53C22001770006) and by the PRIN 2022-PNRR project "MUCES" (CUP E53D23016290001), both funded by the European Union - NextGenerationEU.

References

[1] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, VinVL: Revisiting visual representations in vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[2] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, L. Wang, Scaling Up Vision-Language Pre-training for Image Captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[3] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2018.

[4] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman,

P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, J. Jitsev, LAION-5B: An open large-scale dataset for training next generation image-text models, in: Advances in Neural Information Processing Systems, 2022.

[5] Y. Li, M. Du, R. Song, X. Wang, Y. Wang, A Survey on Fairness in Large Language Models, arXiv preprint arXiv:2308.10149 (2023).

[6] S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, R. Cucchiara, Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models, in: Proceedings of the European Conference on Computer Vision, 2024.

[7] X. Liu, Y. Zhu, Y. Lan, C. Yang, Y. Qiao, Safety of Multimodal Large Language Models on Images and Text, in: Proceedings of the International Joint Conferences on Artificial Intelligence, 2024.

[8] Y. Chen, W. Li, X. Chen, L. V. Gool, Learning Semantic Segmentation From Synthetic Data: A Geometrically Guided Input-Output Adaptation Approach, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[9] M. Fabbri, G. Brasó, G. Maugeri, O. Cetintas, R. Gasparini, A. Ošep, S. Calderara, L. Leal-Taixé, R. Cucchiara, MOTSynth: How Can Synthetic Data Help Pedestrian Detection and Tracking?, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[10] S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, D. J. Fleet, Synthetic Data from Diffusion Models Improves ImageNet Classification, arXiv preprint arXiv:2304.08466 (2023).

[11] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-Shot Text-to-Image Generation, in: Proceedings of the International Conference on Machine Learning, 2021.

[12] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical Text-Conditional Image Generation with CLIP Latents, arXiv preprint arXiv:2204.06125 (2022).

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[14] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, Y. Taigman, Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors, in: Proceedings of the European Conference on Computer Vision, 2022.

[15] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis, arXiv preprint arXiv:2307.01952 (2023).

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common Objects in Context, in: Proceedings of the European Conference on Computer Vision, 2014.

[17] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond Empirical Risk Minimization, in: Proceedings of the International Conference on Learning Representations, 2018.

[18] D. Caffagni, M. Barraco, M. Cornia, L. Baraldi, R. Cucchiara, SynthCap: Augmenting Transformers with Synthetic Data for Image Captioning, in: Proceedings of the International Conference on Image Analysis and Processing, 2023.

[19] R. Socher, L. Fei-Fei, Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2010.

[20] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, S.-C. Zhu, I2T: Image Parsing to Text Description, Proceedings of the IEEE 98 (8) (2010).

[21] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015.

[22] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-Critical Sequence Training for Image Captioning, in: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[23] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: Proceedings of the International Conference on Machine Learning, 2015.

[24] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[25] T. Yao, Y. Pan, Y. Li, T. Mei, Exploring Visual Relationship for Image Captioning, in: Proceedings of the European Conference on Computer Vision, 2018.

[26] X. Yang, K. Tang, H. Zhang, J. Cai, Auto-Encoding Scene Graphs for Image Captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: Proceedings of the International Conference on Learning Representations, 2021.

[29] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: Proceedings of the International Conference on Machine Learning, 2021.

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, in: Proceedings of the International Conference on Machine Learning, 2021.

[31] X. Zhang, X. Sun, Y. Luo, J. Ji, Y. Zhou, Y. Wu, F. Huang, R. Ji, RSTNet: Captioning With Adaptive Attention on Visual and Non-Visual Words, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[32] L. Huang, W. Wang, J. Chen, X.-Y. Wei, Attention on Attention for Image Captioning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[33] Y. Pan, T. Yao, Y. Li, T. Mei, X-Linear Attention Networks for Image Captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[34] W. Liu, S. Chen, L. Guo, X. Zhu, J. Liu, CPTR: Full Transformer Network for Image Captioning, arXiv preprint arXiv:2101.10804 (2021).

[35] S. Herdade, A. Kappeler, K. Boakye, J. Soares, Image Captioning: Transforming Objects into Words, in: Advances in Neural Information Processing Systems, 2019.

[36] M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara, Meshed-Memory Transformer for Image Captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[37] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, R. Ji, Dual-Level Collaborative Transformer for Image Captioning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021.

[38] M. Wu, X. Zhang, X. Sun, Y. Zhou, C. Chen, J. Gu, X. Sun, R. Ji, DIFNet: Boosting Visual Information Flow for Image Captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[39] V.-Q. Nguyen, M. Suganuma, T. Okatani, GRIT: Faster and Better Image captioning Transformer Using Dual Visual Features, in: Proceedings of the European Conference on Computer Vision, 2022.

[40] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, K. Keutzer, How Much Can CLIP Benefit Vision-and-Language Tasks?, in: Proceedings of the International Conference on Learning Representations, 2022.

[41] R. Mokady, A. Hertz, A. H. Bermano, ClipCap: CLIP Prefix for Image Captioning, arXiv preprint arXiv:2111.09734 (2021).

[42] M. Barraco, M. Stefanini, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara, CaMEL: Mean Teacher Learning for Image Captioning, in: Proceedings of the International Conference on Pattern Recognition, 2022.

[43] M. Barraco, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara, With a Little Help from your own Past: Prototypical Memory Networks for Image Captioning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[44] Y. Li, Y. Pan, T. Yao, T. Mei, Comprehending and ordering semantics for image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[45] R. Ramos, B. Martins, D. Elliott, Y. Kementchedjhieva, SmallCap: Lightweight Image Captioning Prompted With Retrieval Augmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[46] Y. Tewel, Y. Shalev, I. Schwartz, L. Wolf, ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[47] N. Rotstein, D. Bensaid, S. Brody, R. Ganz, R. Kimmel, FuseCap: Leveraging Large Language Models for Enriched Fused Image Captions, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2024.

[48] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models, in: Proceedings of the International Conference on Machine Learning, 2023.

[49] D. Bucciarelli, N. Moratelli, M. Cornia, L. Baraldi, R. Cucchiara, Personalizing Multimodal Large Language Models for Image Captioning: an Experimental Analysis, in: Proceedings of the European Conference on Computer Vision Workshops, 2025.

[50] F. Cocchi, N. Moratelli, D. Caffagni, S. Sarto, L. Baraldi, M. Cornia, R. Cucchiara, LLaVA-MORE: A Comparative Study of LLMs and Visual Backbones for Enhanced Visual Instruction Tuning, arXiv preprint arXiv:2503.15621 (2025).

[51] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015.

[52] R. Vedantam, C. Lawrence Zitnick, D. Parikh, CIDEr: Consensus-based Image Description Evaluation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015.

[53] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, Y. Choi, CLIPScore: A Reference-free Evaluation Metric for Image Captioning, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021.

[54] S. Sarto, M. Barraco, M. Cornia, L. Baraldi, R. Cucchiara, Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[55] S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara, BRIDGE: Bridging Gaps in Image Captioning Evaluation with Stronger Visual Cues, in: Proceedings of the European Conference on Computer Vision, 2024.

[56] J. Cho, S. Yoon, A. Kale, F. Dernoncourt, T. Bui, M. Bansal, Fine-grained Image Captioning with CLIP Reward, in: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2022.

[57] N. Moratelli, D. Caffagni, M. Cornia, L. Baraldi, R. Cucchiara, Revisiting Image Captioning Training Paradigm via Direct CLIP-based Optimization, in: Proceedings of the British Machine Vision Conference, 2024.

[58] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, J. Gao, Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks, in: Proceedings of the European Conference on Computer Vision, 2020.

[59] T. DeVries, G. W. Taylor, Improved Regularization of Convolutional Neural Networks with Cutout, arXiv preprint arXiv:1708.04552 (2017).

[60] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random Erasing Data Augmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[61] J.-H. Kim, W. Choo, H. O. Song, Puzzle Mix: Exploiting Saliency and Local Statistics for Optimal Mixup, in: Proceedings of the International Conference on Machine Learning, 2020.

[62] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, B. Lakshminarayanan, AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty, in: Proceedings of the International Conference on Learning Representations, 2020.

[63] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[64] J. Liu, B. Liu, H. Zhou, H. Li, Y. Liu, TokenMix: Rethinking Image Mixing for Data Augmentation in Vision Transformers, in: Proceedings of the European Conference on Computer Vision, 2022.

[65] T. Vu, B. Sun, B. Yuan, A. Ngai, Y. Li, J.-M. Frahm, Supervision Interpolation via LossMix: Generalizing Mixup for Object Detection and Beyond, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024.

[66] M. A. Islam, M. Kowal, K. G. Derpanis, N. D. Bruce, Feature Binding with Category-Dependant MixUp for Semantic Segmentation and Adversarial Robustness, in: Proceedings of the British Machine Vision Conference, 2020.

[67] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, M. Bennamoun, Text to image synthesis for improved image captioning, IEEE Access 9 (2021) 64918–64928.

[68] C. Xiao, S. X. Xu, K. Zhang, Multimodal Data Augmentation for Image Captioning using Diffusion Models, in: Proceedings of the ACM International Conference on Multimedia Workshops, 2023.

[69] W. Li, J. Lotz, C. Qiu, D. Elliott, The Role of Data Curation in Image Captioning, in: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, 2024.

[70] F. Ma, Y. Zhou, F. Rao, Y. Zhang, X. Sun, Image Captioning with Multi-Context Synthetic Data, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024.

[71] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, in: Proceedings of the International Conference on Machine Learning, 2022.

[72] T. Nguyen, S. Y. Gadre, G. Ilharco, S. Oh, L. Schmidt, Improving Multimodal Datasets with Image Captioning, in: Advances in Neural Information Processing Systems, 2023.

[73] D. Hendrycks, K. Gimpel, Gaussian Error Linear Units (GELUs), arXiv preprint arXiv:1606.08415 (2016).

[74] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[75] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer Bormalization, arXiv preprint arXiv:1607.06450 (2016).

[76] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2002.

[77] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops, 2005.

[78] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops, 2004.

[79] P. Anderson, B. Fernando, M. Johnson, S. Gould, SPICE: Semantic Propositional Image Caption Evaluation, in: Proceedings of the European Conference on Computer Vision, 2016.

[80] R. Sennrich, B. Haddow, A. Birch, Neural Machine Translation of Rare Words with Subword Units, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2016.

[81] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: Proceedings of the International Conference on Learning Representations, 2015.

[82] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, H. Wu, Mixed Precision Training, in: Proceedings of the International Conference on Learning Representations, 2018.

[83] S. Rajbhandari, J. Rasley, O. Ruwase, Y. He, ZeRO: Memory optimizations Toward Training Trillion Parameter Models, in: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2020.

[84] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, et al., Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020.

[85] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., DINOv2: Learning Robust Visual Features without Supervision, Transactions on Machine Learning Research Journal (2024) 1–31.

[86] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, J. Jitsev, Reproducible Scaling Laws for Contrastive Language-Image Learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.