Do DALL-E and Flamingo Understand Each Other?

Hang Li[1,2] **Jindong Gu**[3] **Rajat Koner**[1] **Sahand Sharifzadeh**[1] **Volker Tresp**[1,2]
[1]LMU Munich, Germany, [2]Siemens AG, Germany, [3]University of Oxford, UK
hang.li@campus.lmu.de

Abstract

A major goal of multimodal research is to improve machine understanding of images and text. Tasks include image captioning, text-to-image generation, and vision-language representation learning. So far, research has focused on the relationships between images and text. For example, captioning models attempt to understand the semantics of images which are then transformed into text. An important question is: which annotation reflects best a deep understanding of image content? Similarly, given a text, what is the best image that can present the semantics of the text? In this work, we argue that the best text or caption for a given image is the text which would generate the image which is the most similar to that image. Likewise, the best image for a given text is the image that results in the caption which is best aligned with the original text. To this end, we propose a unified framework that includes both a text-to-image generative model and an image-to-text generative model. Extensive experiments validate our approach.

1. Introduction

An important goal of multimodal research is to improve machine understanding of images and text [2, 6, 33, 34, 37]. In particular, considerable effort is centered around the question of how to achieve meaningful communication between both modalities [23, 30, 34, 45, 46]. For example, image captioning should describe the semantic content of an image as a coherent text that can be understood by humans [20, 47]. Conversely, text-to-image generation models [17, 31, 34, 35] utilize the semantics of a textual description for creating a realistic image. This raises interesting questions regarding semantics: For a given image, what textual description most accurately describes the image? Similarly, for a given text, what is the most meaningful image realization? For the first question, some works claim that the best caption for an image should be both natural and informative of the visual content [29, 49]. For the second question, meaningful images should have high quality, diversity, and faithfulness to the text [28, 31]. Motivated by human communication, we argue that the selection of the most accurate image-text pairs could be achieved through an interactive task involving both a text-to-image and an image-to-text based generative model. As shown in Figure 1, in the first task, the image-to-text model is the message sender and the text-to-image model is the message receiver. The goal of the sender is to convey the content of the image to the receiver using natural language such that the receiver can comprehend the language and reconstructs a truthful visual representation. Once the receiver can reconstruct the original information to a high degree of fidelity, the message has been successfully communicated. We argue that the generated textual description is optimal, which results in an image that is most similar to the original image. This principle is inspired by the human use of language for communication. Imagine the following situation. For an emer-

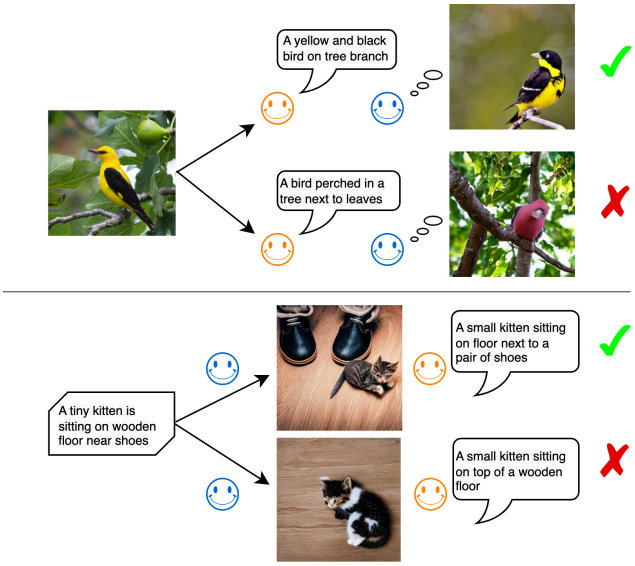

Figure 1. Description of our approach for finding the best caption of an image (above). First, we produce a set of captions for a given image. Then we generate images from those captions. We select the caption for which the reconstructed image is most similar to the original image. Similarly, the task is swapped to generate the best image for a text (below).

gency call, the police are informed by telephone of the circumstances of a car accident and the status of the injured person. This essentially involves the image captioning process of the witness on site. The police need to mentally reconstruct the environment based on the descriptions to organize proper rescue operations. Obviously, the best caption describing the scene should be the best guide for the reconstruction.

The second task involves reconstructing the text: The text-to-image model becomes the sender and the image-to-text model becomes the receiver. Once the two models agree on the content of the information at the textual level, the image used as the medium to convey the message is the best image representing the source text.

Our approach is closely related to inter-agent communication. Language is a major means of exchanging information between agents. But how can we be sure that the first agent has the same understanding of what a cat or a dog is as the second agent? The idea pursued in this paper is to have the first agent analyze an image and produce a text describing that image. The second agent then obtains the text and simulates an image based on the text. This latter process can be thought of as an embodiment process. We propose that communication is successful if the image simulated by the second agent is close to the image the first agent received as input (see Figure 1).

In our experiments, we utilize off-the-shelf models, in particular recently developed large-scale pre-trained models [7, 8, 12, 18, 43]. For example, Flamingo [2] and BLIP [27] are image-captioning models that automatically generate textual descriptions conditioned on images. Likewise, image generation models trained on image-text pairs understand the deep semantics of text and compose high-quality images. An example is DALL-E [34, 35] and the latent diffusion model (SD) [37]. Further, we exploited the CLIP [33] model to compare images or text. CLIP is a vision-language model that aligns image and text representations into a shared embedding space. We use manually curated image-text datasets, such as COCO [9] and NoCaps [1], for evaluating the quality of generated captions. The image and text generation models have stochastic components that allow sampling from the distribution. Thus we can select the best caption or image from a set of candidates. Various sampling methods, including nucleus sampling [19], are proposed for image captioning. We use nucleus sampling as a baseline to show the superiority of our method.

Our contributions can be summarized as:

- We systematically validate our two assumptions concerning the best image for text and the best text for image transformations. We demonstrated that the best caption for an image is the one that leads to the best reconstruction of the original image.
- We show that the best image for a caption is the image that leads to the best reconstruction of the original text.
- We propose an effective way to improve nucleus sampling based on knowledge of text-to-image models.

2. Related Work

Text to Image Generation Text-conditioned image generation models are mainly divided into four groups, namely normalizing flows [36], VAE [25], GAN [14], and recently developed diffusion models [17, 31, 34, 37]. GAN suffers from low training stability due to its adversarial nature. Furthermore, the mode-collapse problem in GAN causes the generated images to be less diverse [41]. A diffusion-based model defines a forward noise process and learns the reverse process. It iteratively samples from the distribution specified by a learned Markov Chain, where a noise vector is gradually denoised into an image [17]. For image generation with text input, the textual information is incorporated in the model using concatenation or cross attention [43] mechanism to guide the generation process. The recently released DALL-E [34] and SD [37] are representatives of such diffusion-based models. However, generating faithful images from the input text is still an open problem. Challenges include reasoning about the semantics between objects and their complex relationships, and binding attributes to corresponding objects [22, 37].

Image Captioning The task of image captioning is to describe a scene using natural language [20]. Due to recent developments in transformer-based large-scale models [27, 48], captions provided by these models are increasingly indistinguishable from those provided by humans. Caption generation models usually consist of an image encoder and a text decoder. The image encoder extracts the visual features of different levels of granularity [4]. These features are passed to the decoder to output coherent sentences. Image captioning is closely related to question-reasoning based on images. Therefore, Flamingo [33] not only achieves the improvement of image captioning but also excels in multimodal reasoning tasks such as visual question answering and text-image retrieval. Unifying visual-linguistic tasks is a current trend due to the shared transformer architecture in the vision and language domains [21].

Vision Language Representation Learning Representation learning for vision and language handles semantic alignment between different modalities [10, 42]. CLIP [33] is trained on large-scale image and text pair datasets to obtain a unified representation of different representations for the same concept. For that, an image encoder and a text encoder separately map images and text into a high-dimensional space, and a distance-based loss is utilized to

enforce representations of identical concepts in neighborhood regions. The challenges of multimodal learning include heterogeneity of different data domains, modal connectivity, and interactive reasoning across domains [5, 15]. In our project, CLIP is used to compute a similarity metric to quantify the distance between instances from the same domain. Extracting features using pre-trained representation models and evaluating the similarity of the original images based on the embedding distance is a widely used technique [10]. Other pre-trained multimodal representation models include, e.g., LXMERT [42] and ALIGN [21].

Visual Grounding Visual grounding is the task of localizing concepts referred to by the language onto an image [11]. By retrieving the concepts mentioned in the language, we can analyze the model's understanding of the concept and gain interpretability. In general, visual grounding is a special case of the embodied cognition paradigm [24]. Unlike methods based on association and statistical data learning, the embodied paradigm emphasizes the importance of learning the meaning of symbols by interacting with the world [40]. Our work explores the problem of text understanding by transformation into image representations, such as the ability to embed visual concepts and relationships into images. Generating images from text is similar to the process of embodied cognition, where the understanding of concepts in the text is displayed in images.

3. Method

Our framework consists of three pre-trained SOTA neural networks. First, an image-to-text generation model, second, a text-to-image generation model, and third, a multimodal representation model consisting of an image encoder and a text encoder to map an image or text into its semantic embeddings, respectively.

3.1. Task Formulation

For a given image $x \in \mathbb{R}^{H \times W \times C}$ of size height H, width W, color channels C, the *captioning task* concerns of generating a description y consisting of words from a vocabulary of natural language. The quality of the caption is evaluated against a set of reference captions $\mathcal{Y}$ using different text distance metrics, see 4.3. The *image generation task* is viewed as the reversed process: For a given textual description y, an image x is generated to present the semantics of the description. The quality of an image can be roughly defined in two ways. The first is to compare the image to realistic images and define high-quality images as the most natural images. The second approach assesses the quality based on the faithfulness of presenting the given semantics. We evaluate the image quality from the second perspective, i.e., the generated image is compared to a set of references image $\mathcal{X}$ using distance metric in the image feature space, see 4.3.

3.2. BLIP: Image Captioning

We use the publicly accessible BLIP [27] model for image captioning in our framework. The BLIP model consists of an image encoder to understand the image features and a text decoder to generate text in an autoregressive manner. The image encoder uses a vision-transformer [13] backbone, which divides an image into a sequence of small patches before processing them with self-attention. Its final outputs are a sequence of embedding vectors $v_i \in \mathbb{R}^d$ which serves as the grounding information for the text generation. Thereafter, a text decoder generates tokens by iteratively attending to previously generated tokens and the encoded visual features. Such interactions are learned through the self- and cross-attention layers in the text decoder. Pre-training of BLIP involves additional components such as a BERT text encoder [12], and uses contrastive loss for text-image matching. After training on weakly coupled image-text data from the web, BLIP is fine-tuned on COCO captioning dataset to enhance its ability for caption generation. Different decoding strategies can be employed to generate sentences, including greedy search, beam search, Top-k sampling, and Top-p sampling. Greedy search always chooses the next word with the highest probability. Beam search selects sentences with maximum probability within a defined beam width. Top-k sampling samples the next word from a set of words that ranks top k among the vocabulary. Top-p sampling, also known as nucleus sampling, generates words from a set of words whose probability sums exceed a threshold probability p. The options to control the sequence length and variations of the sentence are done via parameters in the generation process, including beam width, k value, p value, minimal number of words, the maximal number of words, and repetition penalty.

3.3. SD: Text to Image Generation

We use the latent diffusion model, also known as stable diffusion (SD), to generate images from a text prompt. In SD, the text encoder processes the text to output a sequence of context vectors representing the textual features. The context vector guides the denoising process of noised data from the low-density region to a high-density region in the data space. Hence, image generation in diffusion models is an iterative sampling process. The sampling process begins with a random noise vector that is drawn from, e.g., a uniform Gaussian distribution. At each step, the model improves the vector by estimating the noise in the data and subtracting the noise from the current values. A U-Net [38] is utilized to perform this estimation based on the time embedding vector and the noised vector. Through the cross-attention mechanism, the information of the text is fused to the visual feature vectors. The innovation of SD is that it uses an autoencoder to compress the image into a low-dimensional semantic space and learns the diffusion process

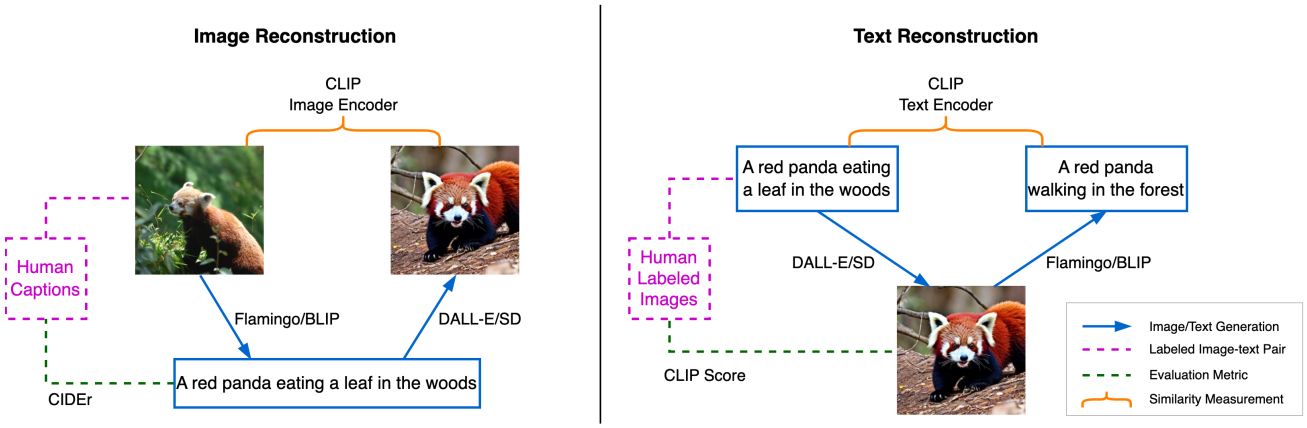

Figure 2. Illustration of our proposed framework. Left: a pipeline for image reconstruction. The input image is fed to Flamingo/BLIP to generate a caption, which is fed to DALL-E/SD to reconstruct an image. The generated image is compared with the input image using the CLIP image encoder in the embedding space. Each input image has human-annotated captions which can be used to evaluate the generated caption. Right: a pipeline for text reconstruction. This shows a similar process, with the input being the text.

in that space. SD is trained on the LAION-400M dataset, a collection of large-scale image-text pairs sourced from the web. The ViT, U-Net, and autoencoder are all frozen in our experiments.

3.4. CLIP: Image and Text Similarity

To evaluate the similarity between the generated image/text and the source image/text, we leverage the semantic space learned by CLIP [33]. CLIP consists of a unimodal image encoder and a unimodal text encoder. For a given image-text pair, CLIP maps them into a shared high-dimensional semantic space. The similarity between the input text and the input image is defined as the cosine distance of their embeddings in this semantic space. During training, the distances between matching image-text pairs are minimized, whereas, mismatched images and text are mapped to different locations in this space. Let $\mathbf{e}_x \in \mathbb{R}^d$ and $\mathbf{e}_y \in \mathbb{R}^d$ denote the embeddings for image x and text y, the cosine similarity between these two embeddings is defined as $f(\mathbf{e}_x, \mathbf{e}_y) = \frac{\mathbf{e}_x \cdot \mathbf{e}_y}{||\mathbf{e}_x|| ||\mathbf{e}_y||}$. The cosine similarity between the embeddings of the source image and the generated image is used to quantify the *mutual understanding* of the image captioning model and the image generation model. When the reconstructed image has a small distance from the source image, we claim both models achieve mutual understanding.

3.5. Image Reconstruction via Text

As shown on the left side of Figure 2, the image reconstruction task is to reconstruct the source image using language as instructions. The effect of this process is to encourage the generation of the optimal caption that describes the source scene. First, the source image x is fed to the BLIP model to generate multiple caption candidates y_k, e.g., *a red panda eating a leaf in the woods*. The set of generated caption candidates is denoted with $\mathcal{C}$. Then the text y_k is fed to the SD to generate an image $x_k^{'}$. Here $x_k^{'}$ refers to the image of the red panda on the ground. Subsequently, CLIP image encoder is used to extract semantic features from the source image and the generated image, $\mathbf{e}_x \in \mathbb{R}^d$ and $\mathbf{e}_{x_k^{'}} \in \mathbb{R}^d$. We then calculate the cosine similarity between these two embedding vectors. The goal is to find the caption candidate y_s with s being the index to the image that is the closest to the source image, i.e., $y_s = \text{argmin}_{y_j \in \mathcal{C}} f(\mathbf{e}_x, \mathbf{e}_{x_j^{'}})$. We evaluate the best caption against the human annotations using the CIDEr metric. As we are interested in the quality of the generated caption, we configure the BLIP model to output captions with approximately the same length. This allows a fair comparison, as more words would convey more information in an image. All models are frozen in this work, without any fine-tuning.

3.6. Text Reconstruction via Image

Figure 2 on the right side shows the reversed procedure of the previous section. The BLIP model needs to guess the source text under the guidance of SD, which has access to the text but can only present its content in image format. The process begins with generating image candidates x_k for the text y using SD. The set of generated image candidates is denoted with $\mathcal{K}$. Image generation with SD involves a random sampling process, where each generation could end up with a different valid image sample in the huge pixel space. This sampling diversity provides us with a pool of candidates to filter out the best images. Next, the BLIP model generates a caption $y_k^{'}$ for each sampled image x_k. Here $y_k^{'}$ refers to the caption *a red panda walking in the forest*. The CLIP text encoder then extracts

features for the source text and the generated text, denoted by $\mathbf{e}_y \in \mathbb{R}^d$ and $\mathbf{e}_{y'_k} \in \mathbb{R}^d$, respectively. The purpose of this task is to find the best image candidate x_s that matches the semantics of the text y. To achieve this, we compare the distance between the generated text and the input text, then select the image whose pairing text has the least distance, i.e., $x_s = \text{argmin}_{x_j \in \mathcal{K}} f(\mathbf{e}_y, \mathbf{e}_{y'_j})$. We argue that x_s can best depict the text description y, since it can convey the content to the receiver with minimal information loss. We treat the image $\tilde{x}$ in the matched pair of text y as the reference presentation of y and quantify the best image as how close it is to the reference image.

4. Experimental Setup

We evaluate our approach on two image captioning datasets. This section describes the details of our models and evaluation metrics.

4.1. Datasets

NoCaps [1] contains large-scale images and captions split into a validation set and a test set with 4,500 images and 10,600 images, respectively. The images are from the Open Images dataset [26]. Each image is annotated with 10 reference captions generated by humans. NoCaps covers 400 object classes and is suited for evaluating image captioning in real-world settings. The validation split is used in our experiments.

MS-COCO Captions [9] consists of 1.5M captions for 330,000 images. The captions are generated by human annotators. Each image is linked to 5 captions. Our experiment uses a randomly sampled subset of 2,000 images from the validation set.

4.2. Models

For image captioning, we use the open implementation of BLIP model[1]. We use the ViT-L/16 for the image encoder and $\text{BERT}_{\text{base}}$ for the text decoder. The model is pre-trained on 129M images from the web and fine-tuned on COCO dataset. For text-to-image generation, we use the pre-trained stable diffusion model[2]. The SD model consists of an Autoencoder with a downsampling factor of 8, 860M U-Net, and CLIP ViT-L/14 text encoder. The model is pre-trained on LAION-5B dataset [39] and fine-tuned for LAION-Aesthetics data. We adopt the pre-trained CLIP model[3] with the ViT-L/14 image encoder. CLIP is trained on 400M text-image pair dataset.

[1]https://github.com/salesforce/BLIP
[2]https://github.com/huggingface/diffusers
[3]https://huggingface.co/docs/transformers/model_doc/clip

4.3. Evaluation Metrics

Image Caption We adopt the commonly used evaluation methods for image captioning and report the captioning score on multiple metrics, including BLEU [32], CIDEr [44], CIDErD [44], and SPICE [3].

Image Generation Although FID [16] is widely used to quantify the quality of the generated images compared to real images, it cannot reveal similarity information between two images at the semantic level. Therefore, we use *CLIP Score* to measure the semantic distance between two images. The CLIP Score between two images x_i and x_j is defined as $\text{CLIP\_Score}(x_i, x_j) = f(\mathbf{e}_{x_i}, \mathbf{e}_{x_j})$, where $\mathbf{e}_{x_i}$ and $\mathbf{e}_{x_j}$ are the embedding vectors obtained from a pre-trained CLIP image encoder.

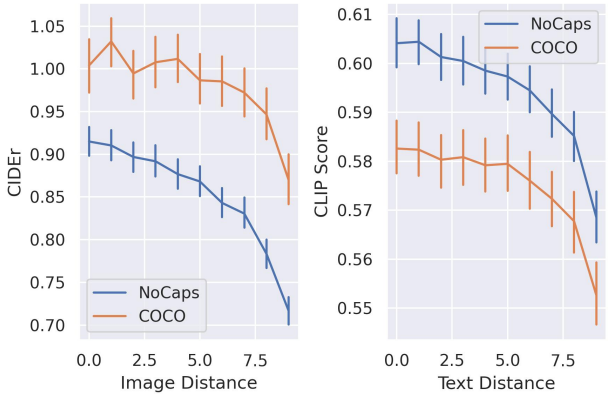

Figure 3. Left: for each given image, the better the reconstructed image (shown in x-axis), the better the caption (shown in y-axis). Right: for each given text, the better the reconstructed text (shown in x-axis), the better the image (shown in y-axis).

5. Evaluation

In this section, we present our experimental results, including the quantitative analysis of the bidirectional transformation tasks demonstrating the superiority of our method compared to the baseline method. Qualitative examples are provided subsequently to demonstrate the characteristics of the proposed method.

5.1. Best Caption for Image

For the image reconstruction task, we first generate ten caption candidates for an image using the nucleus sampling method with $p = 0.9$. We then generate an image for each caption. In the end, we get ten text-image pairs (y_k, x_k) for each input image. The caption y_k is evaluated with the ground truth captions provided in the dataset and the CIDEr score is returned. The CIDEr score quantifies the quality of

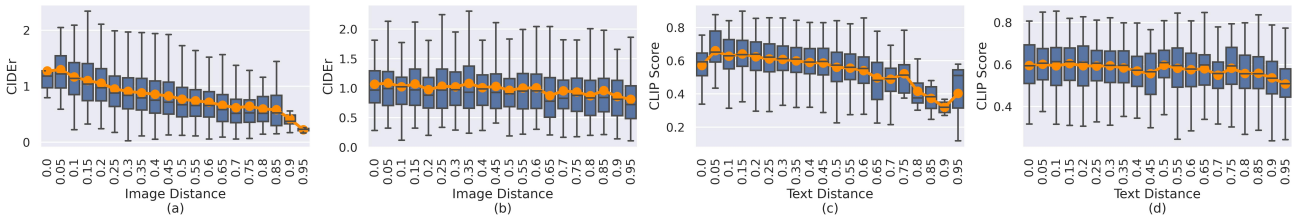

Figure 4. (a) and (b) show a correlation between the quality of generated text (shown in y-axis) and the quality of the image with respect to the source image (shown in x-axis) on NoCaps and COCO dataset, respectively. The scores are averaged for each source image. Likewise, (c) and (d) shows the correlation between the quality of the generated image (shown in y-axis) and the quality of the image (shown in x-axis) on NoCaps and COCO dataset.

	NoCaps							COCO						
Method	B1	B2	B3	B4	CIDEr	CIDErD	SPICE	B1	B2	B3	B4	CIDEr	CIDErD	SPICE
Nucleus	73.0	52.4	36.1	24.1	85.0	74.6	11.6	**66.9**	47.1	32.4	21.9	98.2	90.1	19.6
Ours	**74.3**	**53.8**	**37.3**	**25.2**	**91.5**	**80.3**	**12.3**	**66.9**	**47.2**	**32.5**	**22.0**	**100.4**	**92.0**	**20.1**
Gain (%)	+1.8	+2.7	+3.5	+4.2	+7.6	+7.7	+6.3	+0.0	+0.1	+0.4	+0.3	+2.2	+2.1	+2.2

Table 1. Comparison of nucleus sampling and our proposed method on two datasets. Our method is better than nucleus sampling on any metrics. The relative gain of our method compared to the nucleus sampling is given in the last row.

the generated caption. The image x_k is passed through the CLIP image encoder to obtain an embedding vector. The embedding vectors of all $x_k \in \mathcal{K}$ are compared with the embeddings of the input image using cosine distance. The left image in Figure 3 shows the correlation between the quality of the image reconstruction and the quality of the caption on two datasets. The score is averaged across all images, i.e., we first rank the image distances within the ten pairs for each image. Then we average the y value for each location over all images in the dataset. The variance of the caption scores is shown as a vertical bar. From this figure, we can see that the better the reconstructed image, the better the caption score, i.e., the best text for an image is indeed the text that leads to a better reconstruction. Note that for images with good reconstruction quality, the quality of the caption changes slower than for images with low quality. The reason for this is probably that for good reconstructions there are multiple captions that can interpret the scene, whereas, for bad captions, the reconstruction fails dramatically.

Figure 4 (a) and (b) show the relationship between the image reconstruction and caption quality averaged over images. For each image, the scores of each sampled caption are averaged, as well as the image reconstruction score. We then study the correlation between the captions and reconstruction across images. Again, we find that the better the reconstruction, the better the caption quality. The x-axis is divided into bins of equal length. In each bin, the y-axis shows the statistics of the CIDEr scores of all samples whose image distance falls within this bin. The height of the box shows the boundary of half of the samples for that bin. The line shows the 1.5 quantile value for each bin. The orange curve shows the mean value of CIDEr for each bin. When the text is close enough to the original text, the images also have high reconstruction quality and low variance. These mainly correspond to the images where there is a salient object with distinct features in the center, i.e., the Canadian flag, and butterfly (See Figure 6). Details see discussion in Section 5.4.

5.2. Best Image for Caption

For the text reconstruction task, we generate ten images for each text input. For each image in the validation set, we sample one caption as the source text. Our conclusion is similar to the one in the previous section. Here we use the CLIP text encoder to compare the distance between generated text and the given text. We use CLIP Score to report the quality of the generated image.

The right side of Figure 3 shows the relation between recovered text quality and the generated image quality. Similar to the method used in 5.1, the x-axis shows the quality of the reconstructed text, and the y-axis shows the quality of the reconstructed image. The curve has a declining trend for both datasets, meaning that the best image to depict a text is indeed the image that can best reconstruct the original information. Figure 4 (c) and (d) show the correlation between the text distance and the quality of the reconstructed images. From the figure, we can see that the better the reconstruction, the better the image quality. This conclusion is more obvious in NoCaps than COCO, probably due to the fact that NoCaps contains objects that are not common, which makes the image generation model more uncertain

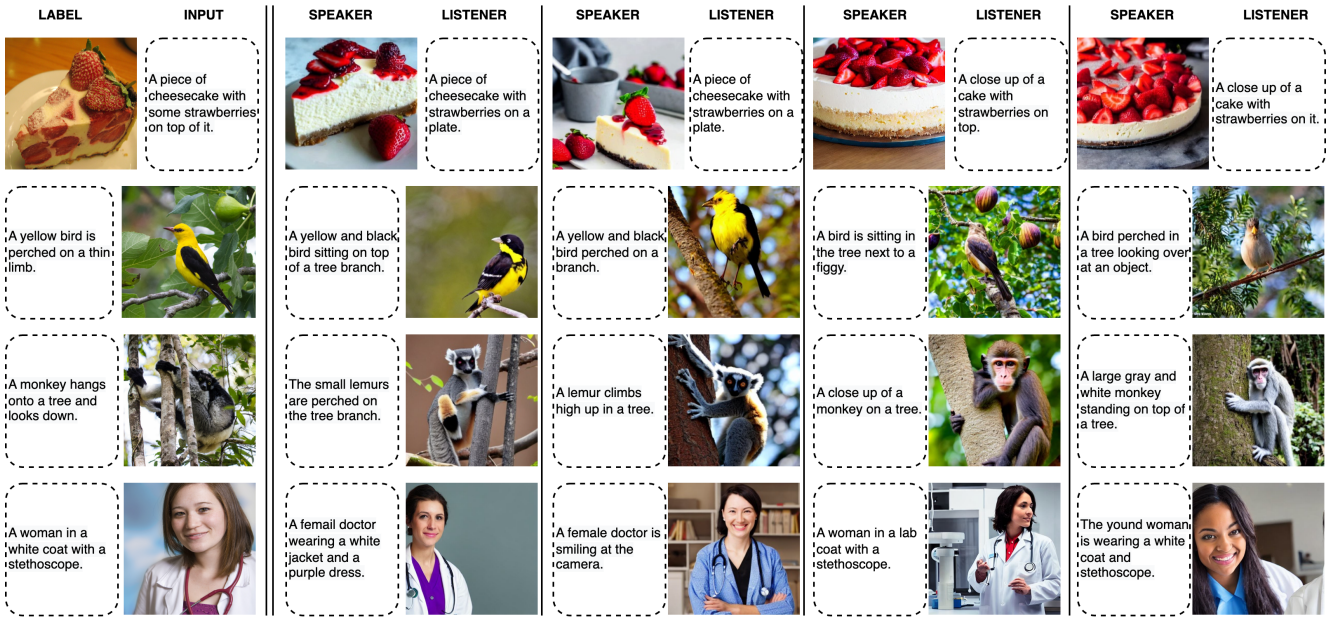

Figure 5. Qualitative examples for both reconstruction tasks. The first two columns in each row show the ground truth image-text pair. The following columns show four generated sample pairs. The second column refers to the input of the model. If the input is text (image), we use the model on the right (left) side in Figure 2. The first row shows four generated images for the input text, together with the captions that describe the images. The generated image-text pairs are sorted by the textual distance. The first two generated captions are close to the input text and the first two images are also the best images for the input text. The remaining three rows show examples from the opposite task which starts with generating captions for images. The third and fourth rows show two negative examples where the closest image does not correspond to the best caption. Details explained in 5.4.3

and difficult to generate the image. Hence we are able to filter out the best image from the diverse set of candidate images. In addition, note that there are variances in each group, highlighting the difficulty of language and image conversion. Concepts in language can have diverse representations in pixel space. Thus, a high variance in the metrics is expected.

5.3. Comparison to Sampling Methods

Nucleus sampling provides a way to sample diverse captions which are close to the maximum likelihood solution. But since there is a trade-off between diversity and performance, it might generate noisy captions. Our proposed framework removes proposed captions that are inaccurate or too noisy.

We run BLIP with ViT-L/16 with nucleus sampling for generating image captions for NoCaps and COCO datasets. The caption generation process of NoCaps/COCO dataset takes 4 hours with an Nvidia T4 GPU. For our method, we generated 10 captions for each image and then an image for each caption, i.e., 45k total RGB images of size of 256×256. The full image generation process took 102 hours on an Nvidia T4 GPU. For text similarity, we then run the CLIP ViT-L/14 image encoder on the original and generated images. We tested with Euclidean distance as well as cosine distance and found no substantial differences. Table 1 shows the superiority of our method over nucleus sampling in every metric. The relative gain of our model can be up to 7.7%.

5.4. Qualitative Results

5.4.1 What is noise in image generation and how can it be eliminated?

To have a better understanding of why our method performs well, we provide some qualitative examples. For example, the first row in Figure 5 describes the input text as *a piece of cheesecake with some strawberries on top of it*. Here we show four generated images and their corresponding text. The first two images have a good caption when running BLIP on them. The generated text is much closer to the original text, thanks to the high-quality reconstruction of the information contained in the text into the visual domain. When these images are fed to BLIP and generate captions again, the information can be well recovered into the text domain due to the strong signals contained in the image. However, the last two images are not aligned with the text, due to the wrong shape of the cake. This can happen because of the stochastic nature of the text-to-image generation process. When such images are fed to BLIP and used

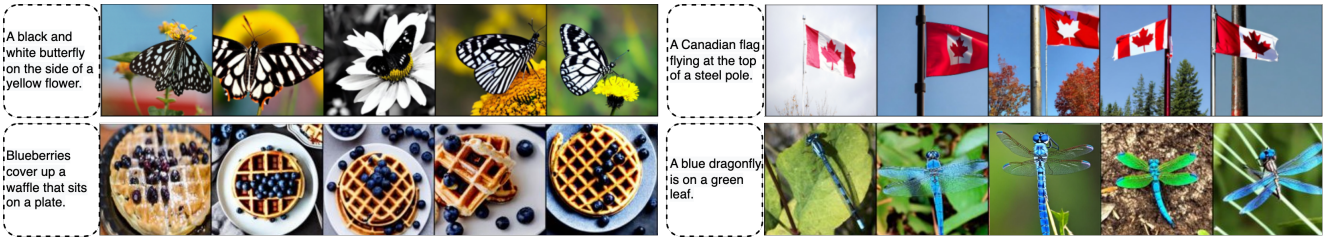

Figure 6. Image samples that show good reconstruction. The first two columns on each side show the ground truth text-image pair. The following four columns show the generated images. These are samples that are easy to reconstruct and have high caption quality.

to generate captions, the captions are much further from the original text. This distance is consistent in the image domain, i.e., the two images are not good images for the text; also they are not close to the oracle image paired for that text.

5.4.2 What is noise in text generation and how can it be eliminated?

Figure 5 (second row) shows an example of the best caption for a given image. On the left, there is an image annotated as *yellow black bird* by human annotators. The BLIP model produces image captions between 10 to 30 words using multiple random seeds. The maximum likelihood training encourages BLIP to generate short sentences, similar to the training data where the average number of words in a caption is 11. Unless there is enough evidence in the image, BLIP will produce short sentences. The diversity of nucleus sampling encourages different captions. Some captions have described the image as *yellow black bird*, whereas others only have *bird* or *small bird*. The captions all seem acceptable to human readers. However, there are some favorable captions that might be better than alternatives. Our work shows a favorable caption is one that leads to good image reconstruction. When the captions are generated back to an image, the four images are different from the original image. CLIP model ranks the first two images as the most similar images. We assume most readers would agree. The reason why the last two images are far from the original one is that the color of the bird is different. This shows that a less significant difference in the text domain might have a large difference in the image domain.

5.4.3 Analyzing Failure Cases

Generation Bias We identify two types of failures. The first type is the biases in the text-to-image generation model and in the image-to-text generation model. Context plays an important role in learning representation. However, some local features might ignore the input but rather follow the context information. For example, the concept of a doctor involves several features such as dress, and equipment. During the training, the model might not disentangle the context of the doctor and its association with specific tools due to their frequent coexistence. Hence the model could wrongly incorporate those factors as being constitutional for the concept of a doctor. This type of bias is prevailing in our analysis, shown in the last row of Figure 5. The generated image is given the input *female doctor wearing a white jacket and a purple dress*. Note that a stethoscope is generated even without being specified in the text. Assumptions are that in the training data, female doctors appear often together with stethoscopes while the text might not contain the word. As a consequence, the caption quality is evaluated as low because of the missing word. However, the image similarity is ranked as high because the salient features, e.g., female, white jacket, purple dress, and stethoscope all match.

Annotation Imperfection Often, human annotators do not annotate concepts in images at a fine-grained level, e.g., due to lack of knowledge. Image captioning models and text-to-image models however have the potential to obtain fine-grained vocabularies, which may outperform those of human annotators. The source image in the third row of figure 5 shows a lemur on a tree branch. Most annotators would probably confuse this species and label it a monkey. But both BLIP and SD have the vocabulary of lemurs so BLIP can recognize the lemur in the image and SD can generate images of lemurs. Hence, the reconstruction is closer to the original image. However, when evaluated against human annotations, it scores lower. This highlights the finding that image captioning models and text-to-image generative models share a vocabulary of concepts, e.g., *lemur*, which more accurately describe objects than a generic concept *monkey*. This brings another advantage of precise description, i.e., *lemur* is more compact than *grey and white monkey*, which apparently needs more bits to convey the information.

6. Conclusion

We have proposed a novel approach to obtain the best matching pairs of text and images by translating both images and text back into their respective domains. Our frame-

work includes an image captioning model, a text-to-image generation model, and two conversational tasks. We investigated the gap between the representations learned by the two models and whether they share the same representation. We found that the two models were able to achieve mutual improvement by interacting with each other, i.e., for images, we find the best captions; and for captions, we find the best images. Our research not only provides insights for improving image and text understanding but also provides a promising direction for the fusion of multimodal models. Future work includes fine-tuning the models to skip the sampling process by generating the best samples directly.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957, 2019. 2, 5

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1, 2

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016. 5

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2

[5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 3

[6] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, pages 1–32, 2021. 1

[7] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 5

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019. 2, 3

[11] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7746–7755, 2018. 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[15] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019. 3

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2

[18] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 2

[19] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019. 2

[20] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019. 1, 2

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2, 3

[22] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018. 2

[23] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27, 2014. 1

[24] Douwe Kiela. Deep embodiment: grounding semantics in perceptual modalities. Technical report, University of Cambridge, Computer Laboratory, 2017. 3

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:1, 2014. 2

[26] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017. 5

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 2, 3

[28] Steven Liu, Tongzhou Wang, David Bau, Jun-Yan Zhu, and Antonio Torralba. Diverse image generation via self-conditioned gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14286–14295, 2020. 1

[29] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 338–354, 2018. 1

[30] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 1

[31] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 4

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1, 2

[36] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 2

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[39] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

[40] Lawrence Shapiro. *Embodied cognition*. Routledge, 2010. 3

[41] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30, 2017. 2

[42] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019. 2, 3

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[44] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5

[45] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 1

[46] Yujia Xie, Luowei Zhou, Xiyang Dai, Lu Yuan, Nguyen Bach, Ce Liu, and Michael Zeng. Visual clues: Bridging vision and language foundations for image paragraph captioning. *arXiv preprint arXiv:2206.01843*, 2022. 1

[47] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 1

[48] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022. 2

[49] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by

distilling image-text matching model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4777–4786, 2020. 1