
Predicting Actions to Help Predict Translations

Zixiu Wu¹ Julia Ive² Josiah Wang¹ Pranava Madhyastha¹ Lucia Specia¹

Abstract

We address the task of text translation on the How2 dataset using a state of the art transformer-based multimodal approach. The question we ask ourselves is whether visual features can support the translation process, in particular, given that this is a dataset extracted from videos, we focus on the translation of actions, which we believe are poorly captured in current static image-text datasets currently used for multimodal translation. For that purpose, we extract different types of action features from the videos and carefully investigate how helpful this visual information is by testing whether it can increase translation quality when used in conjunction with (i) the original text and (ii) the original text where action-related words (or all verbs) are masked out. The latter is a simulation that helps us assess the utility of the image in cases where the text does not provide enough context about the action, or in the presence of noise in the input text.

1. Introduction

Multimodal machine translation (MMT) (Specia et al., 2016) is one of the main applications motivating the creation of the How2 dataset (Sanabria et al., 2018). The goal was to move away from existing datasets – namely Multi30K (Elliott et al., 2016) – with static images and their corresponding simple and short descriptive captions. In the Multi30K dataset, existing work has shown that images can be beneficial, especially in the presence of noisy or incomplete input (Caglayan et al., 2019; Ive et al., 2019).

The language in the How2 dataset is not necessarily descriptive and sentences are longer, less repetitive and structurally more complex. While intuitively this should make the translation task harder and under such conditions one

¹Department of Computing, Imperial College London, United Kingdom ²DCS, Sheffield University, United Kingdom. Correspondence to: Lucia Specia <l.specia@imperial.ac.uk>.

could expect that other modalities could be helpful, the general translation quality obtained by text-only neural machine translation models trained on this dataset is relatively high, as reported in (Sanabria et al., 2018). Additionally, there is not a very close equivalence between the visual and textual modality. For example, many videos are focused on the speaker. Therefore, making use of the additional modality becomes a much harder challenge. As a consequence, previous experiments on MMT on this data thus far have not been able to benefit from images (Sanabria et al., 2018). In this paper we further examine the question of whether visual information can be helpful by (i) using a more advanced model architecture for multimodality, (ii) testing different types of visual features and different ways of representing these features; and (iii) concentrating on the translation of words which we believe the temporal nature of videos could help with.

More specifically, in a similar way to Caglayan et al. (2019), we probe the contribution of images by masking source words to simulate the case of noisy or highly ambiguous input. We focus on actions, which are generally represented by certain verbs, as we believe this is the main additional information one can explore in videos, as compared to static images. We report experiments with a more advanced, transformer-based architecture for MMT than that exploited in Sanabria et al. (2018). Our results show that the visual features, especially those from a CNN fine-tuned for classifying videos into verb-related actions, can be beneficial, in particular for masking settings. Human evaluation of a subset of the data confirms the automatic evaluation results.

2. Dataset and Masking Strategies

We use the How2 (Sanabria et al., 2018) dataset for the experiments, keeping the standard splits:¹ 184,949 training sentences, 2,022 validation sentences and 2,305 test sentences. Our text-only baseline uses the dataset as distributed. For the masking experiments, two strategies to replace words in the source language are defined:

- **Mask action verbs (ACT):** All verbs which correspond to an action as defined in the action categorisations of the Moments in Time dataset (Monfort et al.,

¹<https://github.com/srvk/how2-dataset>

2019) are replaced by a placeholder. The masked words (tokens) make up 2.75%, 2.83%, and 2.84% of the training, validation, and test texts respectively.

- **Mask all verbs (ALL):** All verbs in the sentence are replaced by a placeholder. The masked words (tokens) make up 20.6%, 21.0%, and 20.4% of the training, validation, and test texts respectively.

The masking is performed in all sentences containing (action) verbs in the source language. For that, the data is first POS-tagged and lemmatised using spaCy 2.0.² In the case of action verbs, the resulting lemmatised tokens are matched against the 339 lemmatised action verbs from Monfort et al. (2019).³ The target language remains the same for the purposes of both training and testing. We call the original unmasked sentences *ORG*.

Figure 1 shows some examples of segments from How2 with verbs masked using the two different strategies.

Byte Pair Encoding (BPE) (Sennrich et al., 2015) with 20,000 merge operations is applied on the target training text and each of the differently-masked source training texts separately, leading to 4 distinct vocabularies for *ORG*, *ALL*, *ACT*, and the target language respectively.

3. Visual features

We experiment with three types of visual features:

- **videosum:** the output of the last fully-connected layer of ResNeXt-101 (Xie et al., 2017) with 3D convolutional kernels trained to recognise 400 different actions (Hara et al., 2018);
- **conv4:** the final convolutional layer of a 3D ResNet-50 CNN trained to classify the 339 action verbs from Monfort et al. (2019);
- **emb:** a word embedding matrix for the 339 action verbs, with the embedding of each verb weighted by the final softmax layer of the same CNN for **conv4**.

videosum is provided officially by the How2 Challenge.⁴ Each How2 video segment is divided into 16-frame chunks as separate inputs to the network, according to Sanabria et al. (2018), and the average of the 2048-D feature maps for all the chunks is computed as the single-vector feature of the video segment.

We extract **conv4** and **emb** features using a 3D ResNet-50 CNN trained by Monfort et al. (2019), which inflates a 2D

ResNet-50 CNN pre-trained on ImageNet and fine-tuned on the Moments in Time dataset. We sample 16 equi-distant frames for each video, feed them to the network, and extract the **conv4** and **softmax** vectors from the CNN as the visual features of the video.

For **emb**, we encode each of the 339 category labels as a vector, more specifically a 300-dimensional CBOV word2vec embedding (Mikolov et al., 2013). In the case of multiword phrases, we average the embeddings for each word in the phrase. For each video and for each category label, we scale the category embedding elementwise by its corresponding CNN softmax posterior prediction.

For each video segment in our experiments, **conv4** is represented as a $7 \times 7 \times 2048$ matrix and **emb** as a 339×300 matrix. The former can be interpreted as 49 video region summaries, where each region is a cell of a 7×7 grid that divides the video spatially. The latter can be seen as a description of the video segment based only on the 339 action categories.

4. MMT model

We base our model on the **transformer architecture** (Vaswani et al., 2017) for neural machine translation. Our architecture is a multi-layer encoder-decoder using the **tensor2tensor**⁵ (Vaswani et al., 2018) library.

The encoder and decoder blocks are as follows:

Encoder Block (\mathcal{E}): The encoder block comprises 6 layers, with each containing two sublayers of multi-head self-attention mechanism followed by a fully connected feed forward neural network. We follow the standard implementation and employ residual connections between each layer, as well as layer normalisation. The output of the encoder forms the encoder memory which consists of contextualised representations for each of the source tokens ($M_{\mathcal{E}}$).

Decoder Block (\mathcal{D}): The decoder block also comprises 6 layers. It contains an additional sublayer which performs multi-head attention over the outputs of the encoder block. Specifically, decoding layer d_{i_i} is the result of a) multi-head attention over the outputs of the encoder which in turn is a function of the encoder memory and the outputs from the previous layer: $A_{\mathcal{D} \rightarrow \mathcal{E}} = f(M_{\mathcal{E}}, d_{i_{i-1}})$ where, the keys and values are the encoder outputs and the queries correspond to the decoder input, and b) the multi-head self attention which is a function of the generated outputs from the previous layer: $A_{\mathcal{D}} = f(d_{i_{i-1}})$.

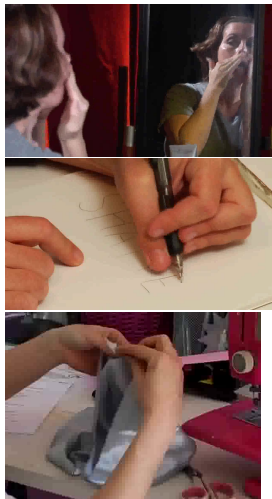
Our multimodal transformer models follow one of the two

⁵<https://github.com/tensorflow/tensor2tensor>

²http://spacy.io/model/en_core_web_lg

³We retain only the verb component for specialised actions such as *playing+music* and *adult+male+singing*

⁴<https://srvk.github.io/how2-challenge/>



- simply apply the cleanser or cream to your hands and apply it to the face and begin rubbing.
 - ◆ simply apply the cleanser or cream to your hands and apply it to the face and begin ▽.
 - ▲ simply ▽ the cleanser or cream to your hands and ▽ it to the face and ▽ ▽.
-
- you can draw it really lightly , go back and erase it later .
 - ◆ you can ▽ it really lightly , go back and erase it later .
 - ▲ you ▽ ▽ it really lightly , ▽ back and ▽ it later .
-
- what we are going to be doing is folding the top over and making a little casing the ribbon will slip through .
 - ◆ what we are going to be doing is ▽ the top over and making a little casing the ribbon will ▽ through .
 - ▲ what we ▽ ▽ to ▽ ▽ ▽ ▽ the top over and ▽ a little ▽ the ribbon ▽ ▽ through .

Figure 1. Three example segments from the How2 training dataset with verbs masked. In each example, the first line (■) shows the full text segment, the second line (◆) shows the segment with verbs from Monfort et al. (2019) masked with ▽, the third line (▲) shows the segment with all verbs masked with ▽.

formulations below for conditioning translations on image information:

- **Additive image conditioning (AIC)** The 2048-D **videosum** feature vector is projected and then added to each of the outputs of the encoder. The projection matrix is jointly learned with the model.
- **Attention over image features (AIF)** The model attends over image features, as in Helcl et al. (2018), where the decoder block now contains an additional cross-attention sub-layer $A_{\mathcal{D} \rightarrow \mathcal{V}}$ which attends to the visual information. The keys and values correspond to the visual information.

For **conv4** and **emb**, the attention is distributed across the 49 video regions and the 339 action categories, respectively. For **videosum**, the 2048-D feature vector is reshaped in row-major order into a 32×64 matrix, so that the attention is over the 32 rows.

Training We keep the hyperparameter settings as in Ivec et al. (2019), i.e. we use the `transformer.big` parameter set with 16 heads, a hidden state size of 1024, a base learning rate of 0.05, and a dropout rate of 0.1 for layer pre- and post-processing at training time. We optimise our models with cross entropy loss and Adam as optimiser (Kingma & Ba, 2014). Training is performed until convergence.⁶ We optimise the number of warmup steps during the multi-GPU training according to Popel & Bojar (2018). We apply beam search of size 10 and alpha of 1.0 for inference.

⁶We use early stopping with patience of 10 epochs based on the validation BLEU score.

Table 1. Results for the test set. We report BLEU scores. Bold highlights our best results.

SETUP	ORG	ACT	ALL
text-only	55.9	53.6	44.1
AIC-VIDEOSUM	55.6	53.6	44.2
AIF-VIDEOSUM	55.7	53.3	44.0
AIF-CONV4	55.6	53.8	44.4
AIF-EMB	56.2	53.5	44.5

Table 2. BLEU deltas recorded when models are given incongruent visual features, compared to the scores reported in Table 1

SETUP	ORG	ACT	ALL
AIC-VIDEOSUM	↑ 0.1	↓ 0.7	↓ 1.0
AIF-VIDEOSUM	↓ 0.1	↓ 0.3	↓ 0.5
AIF-CONV4	↓ 0.3	↓ 0.5	↓ 0.8
AIF-EMB	↓ 0.1	↓ 0.4	↓ 0.3

5. Results

Table 1 reports the results⁷ of our experiments using BLEU (Papineni et al., 2002) as metric.⁸

Our unmasked TEXT-ONLY baseline achieves a BLEU score of 55.9. As expected, the scores for masked models are

⁷The scores may not be comparable to the ones on the How2 Machine Translation Challenge leaderboard, since we do not remove punctuation from the model inferences, but it is the case with submissions to the challenge. The same caveat applies to Table 2.

⁸We measure the performance with Multeval (Clark et al., 2011). We use tokenised and lowercased reference and hypotheses, as in the official challenge.

Predicting Actions to Help Predict Translations



EN	So, how do I make sure that I <u>spin</u> all the way around, or, how do you make sure?
text-only	Então, como eu me certifico de <u>cortar</u> a toda a volta, ou como você se certifica?
AIF-conv4	Então, como eu me certifico de <u>dar a volta</u> , ou, como você se certifica?
AIF-emb	Então, como eu me certifico de <u>viro</u> todo o caminho, ou, como você se certifica?
PT	Então, como eu me certifico de <u>girar</u> ao redor, ou, como você se certifica?

(a) AIF-conv4 guesses the masked word spin correctly as dar a volta, while the text-only model translates it incorrectly as cortar (cut) and AIF-emb translates it partially correctly as virar (turn)



EN	In this clip we're talking about footwork, we're going to be <u>covering</u> the moving forward aspect of it.
text-only	Neste clipe, estamos <u>falando</u> de trabalho de pés, vamos <u>discutir</u> o aspecto da frente dele.
AIF-conv4	Neste clipe, estamos <u>falando</u> de trabalho de pés, vamos <u>cobrir</u> o aspecto da mudança para frente.
AIF-emb	Neste clipe, estamos <u>falando</u> de footwork, vamos <u>discutir</u> o aspecto de movimento em movimento.
PT	Neste vídeo, estamos <u>falando</u> de trabalho de pés, vamos estar <u>cobrindo</u> o aspecto de avançar.

(b) AIF-conv4 guesses talking and covering correctly as falar (talk) and cobrir (cover); the other models get the first word right, but translate the other word as discutir (discuss)



EN	I might use the sixty degree wedge a bit too, but the sand wedge obviously is useful for getting out of the ruff, <u>hitting</u> the ball from the fairway, getting out of sand.
text-only	Eu poderia usar a cunha de sessenta graus também, mas a cunha de areia, obviamente, é útil para sair do ruff, <u>tirar</u> a bola do fairway, sair da areia.
AIF-conv4	Eu poderia usar a cunha de sessenta graus um pouco também, mas a cunha de areia obviamente é útil para sair do pescoço, <u>bater</u> na bola do fairway, saindo da areia.
AIF-emb	nós Eu poderia usar a cunha de sessenta graus um pouco também, mas a cunha de areia obviamente é útil para sair do ruff, <u>bater</u> a bola do fairway, sair da areia.

(c) AIF-conv4 and AIF-emb guess hit correctly as bater, while the text-only model translates it as tirar (remove)

Figure 2. Examples of improvements of AIF-conv4 and AIF-emb over the text-only baseline. Underlined text denotes masked words and their translations.

lower: for ACT, we observe a BLEU of 53.6; for ALL a BLEU of 44.1.

Overall, AIC-videosum is on par with TEXT-ONLY for ACT and ALL. For ORG, a drop of 0.3 BLEU is recorded.

AIF-videosum, on the other hand, leads to degraded performance: a 0.2 decrease for ORG, 0.3 for ACT, and 0.1 for ALL. This reveals the possibility that a single global feature vector is better used as a whole as opposed to as segmented sub-vectors for attention.

conv4 and **emb** features contribute to our best-performing models for different masking settings. AIF-conv4 achieves deltas of -0.3, 0.2 and 0.3 points over the text-only model respectively for ORG, ACT and ALL. The figures for AIF-emb models are 0.3, -0.1 and 0.4 BLEU points on the

same settings. The above also means that AIF-conv4 is our best for ACT and AIF-emb takes the first place for ORG and ALL.

Overall, the top-performing models achieve modest improvements over the text-only baseline in each masking setting. To probe into how much visual features matter to the multimodal models, we conduct incongruent decoding Caglayan et al. (2019), where we feed the visual features in reverse order. The assumption is that models that have learned to exploit visual information to aid translation will suffer from the mismatch and hence result in performance deterioration. We summarise the results of our incongruent decoding for the multimodal models in Table 2.

As expected, visual incongruence leads to worse results (as

much as a 1.0 BLEU drop) in almost all the settings, which proves that multimodality indeed exerts positive influence on the translation. Also evident from the table is that the score changes are considerably more pronounced for ACT and ALL than for ORG, pointing to the possibility that the verb masking leads to the visual features being more relied upon for translation and that a mask-free source sentence may be sufficient already for quality translation.

To sum up, a major finding is the importance of the visual features to the multimodal translation despite the generally moderate improvements achieved in terms of the automatic metric. Surprisingly, the delta in BLEU between text-only and multimodal models for the masked datasets is not substantially larger than for the original dataset (0.3 BLEU points). It is smaller for ACT (0.2 BLEU points), and similar for ALL (0.4 BLEU points).

Additionally, we note that in none of the model settings are the video features able to help bridge the gap between ORG and ACT performances, not even in AIF-conv4 and AIF-emb where the visual features are from a CNN fine-tuned for classifying videos into classes whose labels are closely related to the verbs masked in ACT. However, the gap between ORG and ALL is slightly smaller for some multimodal models than for the text-only model.

5.1. Human Analysis

Automatic metrics often fail to capture nuances in translation quality, such as the ones we expect the visual modality to help with, which – according to human perception – lead to better translations (Elliott et al., 2017; Barrault et al., 2018). We thus performed human evaluation of our best outputs involving native speakers of Portuguese (four annotators) who are fluent speakers of English. We focused only on the evaluation for ACT.

The annotators were asked to rank randomly selected test samples according to how well they convey the meaning of the source (50 samples per annotator). For each source segment, the annotator was shown the outputs of three systems: text-only, AIF-conv4 and AIF-emb. They also had access to reference translations. A rank could be assigned from 1 to 3, allowing ties (Bojar et al., 2017). Annotators could assign zero rank to all translations if they were judged incomprehensible.

Following the common practice in human evaluation for many machine translation shared tasks (Bojar et al., 2017), each system was then assigned a score which reflects the proportion of times it was judged to be better than or equal to the other two systems.

Table 3 shows the human evaluation results. Contrary to the automatic evaluation results for ACT, the AIF-emb setup is generally favoured by human preference. text-only

and AIF-conv4 demonstrate similar performance.

Figure 2 illustrates some cases where AIF-emb or AIF-conv4 outperforms the text-only model.

text-only	AIF-conv4	AIF-emb
0.75	0.73	0.81

Table 3. Human ranking results for ACT: micro-averaged rank over three annotators.

6. Conclusions

We investigated a state of the art multimodal machine translation approach on the How2 dataset. Our focus was on exploring visual features that attempt to represent action information, and on probing their contribution when the input text is corrupted to remove action-related words. The hypothesis was that a well designed multimodal model based on informative visual features should be able to recover from the lack of textual information by leveraging the visual information. Our main results are as follows: (i) performance improvements over the text-only baseline can be achieved by the multimodal models, and the best in all three masking settings are produced by the models that exploit visual features from an action classification network; (ii) visual information is important to the multimodal models, especially in the verb-masked cases where substantial performance drops are observed in case of a visual feature mismatch; (iii) human evaluation indicates that representing features of verb-related actions with word embeddings to exploit similarities between respective verbs could be beneficial. These are promising results for multimodal machine translation and for the use of action-related visual features in this context.

Acknowledgement

This work was supported by the MultiMT (H2020 ERC Starting Grant No. 678017) and MMVC (Newton Fund Institutional Links Grant, ID 352343575) projects. We also thank the annotators for their valuable help.

References

- Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., and Frank, S. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 304–323. Association for Computational Linguistics, 2018. URL <http://aclweb.org/anthology/W18-6402>.
- Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Lo-

- gacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pp. 169–214. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-4717. URL <http://aclweb.org/anthology/W17-4717>.
- Caglayan, O., Madhyastha, P., Specia, L., and Barrault, L. Probing the need for visual context in multimodal machine translation. *CoRR*, abs/1903.08678, 2019. URL <http://arxiv.org/abs/1903.08678>.
- Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 176–181, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/P11-2031>.
- Elliott, D., Frank, S., Sima'an, K., and Specia, L. Multi30k: Multilingual english-german image descriptions. In *5th Workshop on Vision and Language*, pp. 70–74, Berlin, Germany, 2016. URL <http://aclweb.org/anthology/W16-3210>.
- Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pp. 215–233, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W17-4718>.
- Hara, K., Kataoka, H., and Satoh, Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, 2018.
- Helcl, J., Libovický, J., and Varis, D. CUNI system for the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 616–623. Association for Computational Linguistics, 2018. URL <http://aclweb.org/anthology/W18-6441>.
- Ive, J., Madhyastha, Swaroop, P., and Specia, L. Distilling Translations with Visual Awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y. (eds.), *Proceedings of the 1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, Scottsdale, AZ, USA, May 2013. URL <http://arxiv.org/abs/1301.3781>.
- Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., Gutfrueud, D., Vondrick, C., et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–8, 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2019.2901464.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. URL <http://www.aclweb.org/anthology/P02-1040>.
- Popel, M. and Bojar, O. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70, 2018.
- Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., and Metze, F. How2: A large-scale dataset for multimodal language understanding. *CoRR*, abs/1811.00347, 2018. URL <http://arxiv.org/abs/1811.00347>.
- Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.
- Specia, L., Frank, S., Sima'an, K., and Elliott, D. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation*, pp. 543–553. Association for Computational Linguistics, 2016. doi: 10.18653/v1/W16-2346. URL <http://www.aclweb.org/anthology/W16-2346>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.
- Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., et al. Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416*, 2018.
- Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.