
CNN Assistance in Jigsaw Puzzle Solver

Tianbao Li

Department of Computer Science
University of Toronto
Toronto, ON M5S 1A1
tianbao@cs.toronto.edu

Abstract

This paper presents an innovative way to assist solving the jigsaw puzzle with the help of deep convolutional neural networks. The main goal of this project is to predict whether two pieces of the jigsaw puzzle should be neighbors in the original image. Proceeding from traditional methods of using color (like RGB) distance, here, we introduce the low-level features in the image which are extracted by convolutional neural networks. Compared with color-based solutions, using the feature maps generated can help with a deeper intuition on the correlation between edges. The proposed algorithm can achieve considerable accuracy on adjacency prediction.

1 Introduction

Jigsaw puzzles were first introduced around 1760 for map research and then became a popular intelligence entertainment [4] in the last few centuries. The original image is divided into $N \times M$. To solve this, people need to cluster similar pieces, find the neighbors and then reconstruct the original image. However, due to the possible locations and relationship behind pieces, there are quite a huge number of solutions ever for a small jigsaw puzzle, for example, $(8 * 8)! = 1.2688693 * 10^{89}$ solutions for an $8 * 8$ puzzle. Indeed, this problem has been proven to be an NP-complete one [1, 2]. Though tough, actually, jigsaw puzzle solving is quite meaningful beyond the intellectual challenge. It helps a lot in combining shredded documents [9, 10] and even recovering artifacts debris[7].

Recently year, there have been a lot of researchers working on this problem and achieved much progress. However, one of the most important sub-problems in this procedure, determining adjacency between pieces, seems popular for long. When people solve the puzzle manually, they need image information at edges, such as color, texture, instance to make the judgment whether two pieces are neighbors. However, most of the previous research only eyes on the color relationship such as RGB distance. This makes such solutions quite bad for simple-colored images, especially in reconstructing printed documents or ancient artifacts.

Here, we want to mimic how people really solve jigsaw puzzles and work on finding features capable to help make the adjacent prediction. Recently, Convolution Neural Networks (CNN) leads the research area of computer vision with its capability of extracting inside features underneath the images and solve hard problems such as detection[14] and segmentation[5, 19]. Based on the existing tremendous structures, people start to apply transfer learning and use extracted features to solve related computer vision problems[13].

In this paper, we contribute in providing neural networks to predict whether two pieces from a jigsaw puzzle are adjacent in the original image. The neural networks judge the jigsaw piece pair by extracted low-level image features from pre-trained CNN together with color information at edges and output whether these two pieces are neighbors. This model achieves outstanding prediction accuracy in images from ILSVRC2012 dataset[15].

2 Related Work

Jigsaw has been researched on for many years. One most basic idea is to evaluate the compatibility of the adjacent pieces and take a strategy, such as greedy search, to arrange the pieces. One famous work is the Genetic Algorithm (GA) [16]. Given initial candidate solutions, it applies operations like selection, reproduction, and mutation based on the color-distance fitness function. Similarly but innovative, [17] first introduces deep neural network to jigsaw solver and transforms the puzzle into the piece pair adjacency prediction. It samples piece edges to learn the adjacent likelihood through DNN based on the color distance. For these two solvers, they only use color information to judge whether two pieces should be together. However, human use some others like texture to solve. Also, though high accuracy, these solutions could not solve the jigsaw puzzle which generated by single-colored or colorless images. So, there are still some further steps on it.

Recently, with the prosper of convolutional neural networks (CNN) in computer vision area, convolutional neural networks also becomes a good tool to solve jigsaw puzzles. Because of the fact that a good CNN is hard to train and the training data needs many images and manual labels, more and more researchers choose the apply transfer learning on pre-trained models, fine turning on the last a few layers or using it as a feature extractor, instead of going over all the training process [13]. CNN can extract regional features in many perspectives, such as color, texture, pattern, instance, etc. Features in the formers usually reflect the basic information of the images, while the latter layers can make a high-level judgment. These can be good inference for judging adjacent pieces. So, [3, 12] choose to use feature maps from pre-trained CFN [12] (siamese-enned AlexNet [8]), VGG [6] or Resnet [18] and predict the location. However, the main problem for these approaches is that they hold a siamese structure with shared weights for each location, which means they can only solve a limited number for pieces (3×3 in [12], 2×2 and 2×3 in [3]). With the piece amount increasing, the network becomes extremely heavy to train.

3 Preliminary

3.1 Problem Definition

The jigsaw puzzle problem aims to reshape multiple non-overlapping pieces segmented from the original image to the correct arrangement. In our project, like in most computer research, we focus on equal-sized squared $W = N \times M$ pieces. It refuses the aid of shape matching but asks for more on image information. The expected result is a set of indexes $I = (p_1, p_2, \dots, p_W)$ which p_i is the corresponding index for each piece in the original image. For the worst case, it takes the complicity of $O(W!)$. For example, Figure 1 is a 3×3 jigsaw puzzle. The left subgraph is the original image, and the right subgraph is the puzzle which $I = \{6, 2, 7, 3, 5, 4, 0, 8, 1\}$.

3.2 Motivation

Starting from the idea in [17], instead of solving the jigsaw puzzle from the beginning to the end, we choose to focus on the prediction whether two jigsaw puzzle piece should be neighbors in the original image. In recent research, nearly all of the researchers decide to use color information to judge whether two pieces are adjacent. And as in the result, the color distance like RGB distance appears as a good measurement. In some work[16] the color distance is defined as the sum of squared color differences in three-dimensional color space like RGB. For image piece x_i and x_j as a $K \times K \times 3$ matrix which K is the edge length and cb is the color band, the distance is

$$D(x_i, x_j, r) = \sqrt{\sum_{k=1}^K \sum_{cb=1}^3 (x_i(k, K, cb) - x_j(k, 1, cb))^2}$$

However, this color-based measurement doesn't work all the time. For example, in Figure 2, RGB-based solutions cannot predict the similarity of grey-scaled images. Also even changed it to grey-level distance, it doesn't work so well. Also, like in Figure 3, even in the RGB band, the main part of the image is in color blue, even all blue pixels in some edges. So the distances are so close for all edge pairs, which can lead to high error rate for the prediction. Moreover, for the images of furs, hairs and others with obvious pattern structure, like in Figure 4, it is easy to have a low distance with the



Figure 1: Colorful Jigsaw Puzzle Example



Figure 2: Black Jigsaw Puzzle Example



Figure 3: Blue Jigsaw Puzzle Example

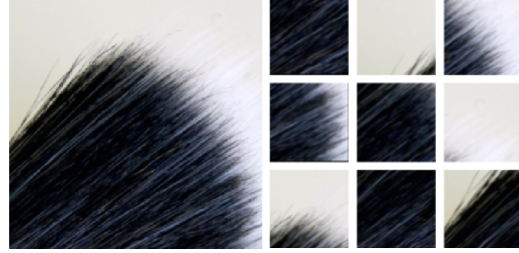


Figure 4: Textural Jigsaw Puzzle Example

displacement fitting the pattern though it is wrong. Only color information can not have a good result on this.

Here, we combine the features extracted from Resnet34[6]. Usually, people choose to extract the feature maps after the activation functions which contains much image information such as color, texture and so on. Resnet34 is constructed on the structure of BasicBlock (several layers of convolutions, batch normalization, ReLUs, etc) and can skip some of the blocks to achieve the residual learning with more blocks. Resnet34 has 3, 4, 6 and 3 blocks in each section. There are some discussions in [12] that feature maps from low layers of CNN can help with solving the jigsaw puzzle better. In our algorithm, we choose to use the ReLU output after the first block.

Combining the measurements above, the goal of the algorithm is to build the neural network based classifier to predict the adjacency likelihood of a pair of two jigsaw puzzle edges in the original image. Note that not all pixels help with the prediction near the edge, we only use the feature maps and color values for the two rows near the edge to reduce the workload of the neural networks.

4 Adjacency Prediction Algorithms

4.1 BuddyNet

In this section, we will discuss the neural network structure, BuddyNet. It is an NN classifier which predicts the adjacency likelihood of a pair of jigsaw pieces in the original image.

As mentioned in the motivation, we want to extract low-level features in Resnet34[6]. Resnet34 has the structure of a combination of basic blocks as in Figure 5. To extract features for each piece, we first resize it into the shape of 224×224 , which is the expected input size of Resnet34. After one convolutional layer of 7×7 , the data is passed through the first basic block and output 64 feature maps which have the size of 56×56 . One example is shown as in Figure 6. 64 small images on the right side are the feature maps of the piece on the left. We can see that some feature maps focus on different regions (the front object or the background), together with some others with more concentration on the shape and other content. To build the connection near the edges, we extract the two rows of pixels near the edge. In this way, we have $2 \times 56 \times 64 = 7168$ values to express the edge.

Also, as an import indicator of the image, we integrate the above features with RGB value. With each pixel in the piece has 3 color band, we have $2 \times 224 \times 3 = 1344$ values (2 rows with the edge of 224 pixels).

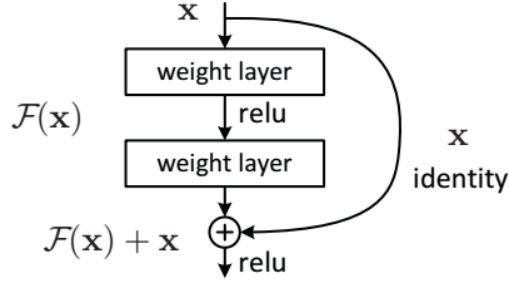


Figure 5: Basic Block of Resnet34

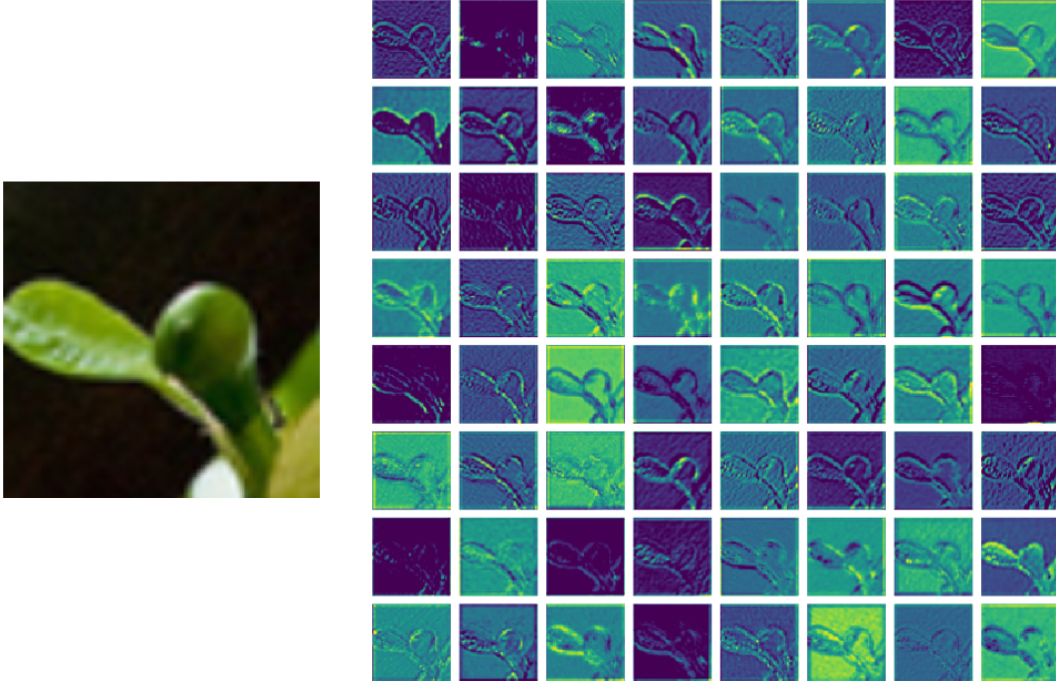


Figure 6: Feature Map Example

To predict the adjacent likelihood, we integrate the information above and build a forward neural network. For a pair of piece edges, we transform the information into a vector with the size of $2 * (7168 + 1344) = 17024$ as the input. The network has 8 fully connected layers of size 8512, 8512, 2464, 2464, 672, 672, and 2. The output is a layer of 2 nodes generated by ReLU activation function. The expected output is (1,0) for neighbor pair (positive) and (0,1) for non-neighbor pair (negative). All the activation functions in the network are ReLU. The structure of the network is shown as Figure 7, with neuron amounts at the bottom.

Additionally, to test how our proposed neural network works for grey-scaled images, we change the architecture into a new version without integrating RGB information. This network takes a vector with the size of $2 * 7168 = 14336$ as the input, and hidden layers have 7168, 7168, 1792, 1792, 448, 448, and 2 neurons. Other parameters are just the same as previous BuddyNet.

4.2 RGBNet

As used in many popular solutions, RGB distance has a fairly good result in neighbor piece prediction. Here, to test the capability of the image expression, we also build a forward neural network only using RGB values. The structure is similar to the one in [17], but we make some revision to achieve a

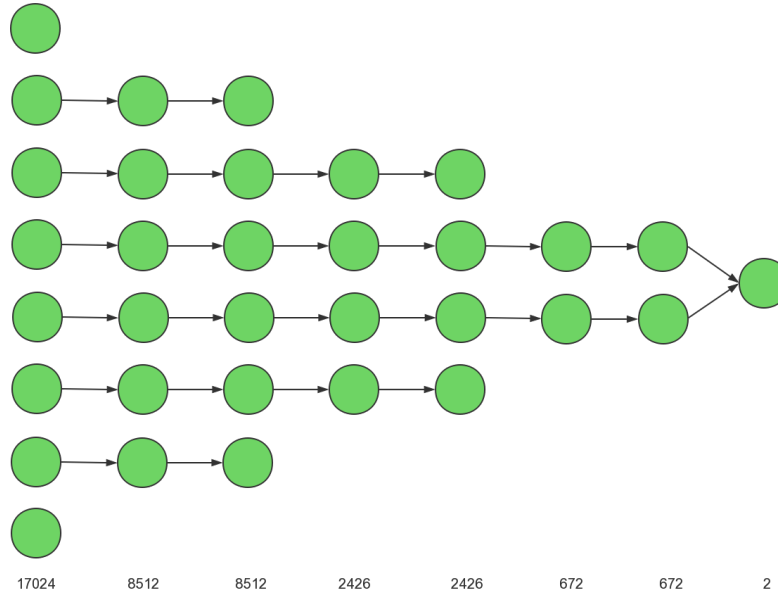


Figure 7: Architecture of BuddyNet

higher accuracy. This RGBNet has 6 fully connected layers, which the amounts of neurons are 2688, 1344, 672, 168, 42, and 2.

5 Evaluation

5.1 Pre-trained Model and Dataset

Instead of training Resnet34 for a huge amount of time, here, we choose to use the pre-trained in the version of PyTorch¹. Other neural networks are also implemented by PyTorch². The Resnet34 models above is pre-trained on ImageNet[8] to learn the general expression of images. To make the evaluation result more persuasive, we separate ImageNet validation set into training and validation set for this project.

The ImageNet validation set has 51003 images in total. The image sizes differ, so we apply random crop on them to 224×224 . Next, each image is divided into pieces. In this evaluation, we try on 3×3 with the edge length of 75 pixels and 8×8 with the edge length of 28 pixels. Then we pass pieces through Resnet34, get the feature maps and generate the vector mentioned above together with RGB bands.

5.2 Data Sampling

For a jigsaw puzzle with N pieces each edge, there are $N * N$ pieces in total and $(N * N)!$ possible combinations of piece pairs. However, only $2 * N * (N - 1)$ among them hold the positive label (actually adjacent in the original image). Note that we do not consider rotated pieces by now. For the algorithm training a classifier, it is extremely important to sample randomly and uniformly. If using the data above, classifying by giving the prediction of “false” can get a high accuracy but low recall, not the classifier we want. To reduce the influence caused by data inclining, we only sample $2 * N * (N - 1)$ pairs from the negative-labeled set (not adjacent ones) to keep a balanced dataset whose size is $4 * N * (N - 1)$.

¹<https://download.pytorch.org/models/resnet34-333f7ec4.pth>

²<https://www.pytorch.org/>

Table 1: Experiment Results on Different Algorithms

	data amount	pieces	training accuracy	validation accuracy
BuddyNet	10000	3×3	0.9546	0.9395
RGBNet	10000	3×3	0.9904	0.9861
DNN-Buddies [17]	10000	3×3	0.9858	0.9846

Table 2: Experiment Results on Different Training Data Sizes

	data amount	pieces	training accuracy	validation accuracy
BuddyNet	10000	3×3	0.9546	0.9395
BuddyNet	100000	3×3	0.9912	0.9749
RGBNet	10000	3×3	0.9904	0.9861
RGBNet	100000	3×3	0.9976	0.9947

In the end, we obtain balanced datasets of 100000 pairs for 3×3 puzzle and 10000 pairs for 8×8 puzzle. We train the forward neural networks with Stochastic Gradient Descent with a momentum of 0.9 to minimize the cross-entropy loss for 200 iterations and apply dropout (ratio = 0.5) on BuddyNet. We use an initial learning rate of 0.1 and it is decayed by a factor of 0.3 every 25 epochs. The default batch size is 32.

5.3 Experiment Results

In this section, we will give out the experiment result on changing different parameters, and analyze how the algorithms work and why is like that.

algorithm

The comparison of different algorithms is shown in Table 1. From the table, we can see that the algorithms only based on RGB bands surpasses our proposed BuddyNet with a margin. And our revised one has a little higher accuracy than DNN_Buddies[17].

Actually, there are only 4916 neurons in RGBNet architecture but 23298 in the proposed BuddyNet. BuddyNet structure is much more complicated than those and requires more training data to overcome underfitting. Also, we can see that there is still a large margin between training accuracy and validation accuracy for BuddyNet, more data can help.

data amount

The comparison of different training data sizes is shown in Table 2. From the table, we can see that more training data helps the models achieve a better result on accuracy, especially for BuddyNet with a complicated structure as discussed above. Also, the existing margin between training accuracy and validation shows there is still potential for BuddyNet.

pieces

The comparison of different piece amounts of each edge is shown in Table 3. From the table, we can see that prediction on 3×3 pieces (edge length of 75) is better than 8×8 (edge length of 28) from the same size of the original image. This is easy to understand because the feature maps and RGB bands we use are all at the pixel level. Fewer pixels mean less information we have, leading to worse prediction. So, for the low-resolution pieces, applying super-resolution methods before passing through Resnet34 may help work better.

color

Table 3: Experiment Results on Different piece amounts

	data amount	pieces	training accuracy	validation accuracy
BuddyNet	10000	3×3	0.9546	0.9395
RGBNet	10000	3×3	0.9904	0.9861
BuddyNet	10000	8×8	0.9277	0.8945
RGBNet	10000	8×8	0.9488	0.9301

Table 4: Experiment Results on Grey-scaled Images

	data amount	pieces	training accuracy	validation accuracy
BuddyNet_grey	10000	3×3	0.8142	0.7693
BuddyNet_grey	100000	3×3	0.8430	0.8085

The result of grey-scaled BuddyNet is shown in Table 4. We use BuddyNet which only accepts feature maps from Resnet34. From the table, we can see that the accuracy is fairly good since grey-scaled jigsaw puzzle is impossible for RGB-based solutions.

6 Additional Trials and Future Work

Besides the achievements above, during this project, we also attempt to contribute and evaluate in some other aspects. These points out where the potential research is going:

- We attempt to build an end-to-end jigsaw puzzle solver using reinforcement learning methods such as deep Q-learning. We try to transform the outputs of BuddyNet and RGBNet into a float value of likelihood and use it as the reward for each operation in the policy search. However, the main problem is that randomly sampling among the available pieces and learning the policy by keeping trying is a high computational cost. When there are many pieces from one puzzle, this kind of solution seems impossible. One idea here is to do clustering among all edges and build an edge-graph using the relationship of same piece edges and neighbor edges. Some graph algorithm like spanning tree or minimum cutting could help. Also, traditional solutions use greedy searching or genetic algorithm and try to improve the initial proposal in several epochs. The classifiers we proposed can also be integrated into these algorithms to improve end-to-end solvers.
- As mentioned in [11], we make some exploration on the batch size. It seems true that training by the smaller size of batches (like 32) can get better result easier and more quickly. Trying to use batches of 1024 or 2048 items, we find that it keeps on pretty low validation accuracy (55% to 60%) for many epochs. And it usually converges at a local minimum with a bad result. More digging on training batch sizes cannot only achieve better validation accuracy but also understand the mechanism behind the black boxes of deep learning.
- Our proposed methods, by now, can not achieve a pretty high prediction accuracy for smaller pieces of jigsaw puzzles. Fewer pixels and less information on each edge hurt the result much a lot. Some more research could be on whether the puzzle-piece super-resolution can help on this.
- Grey-scaled jigsaw has a lot of use cases in real work, including but not limited to archaeology, literature and criminal cases. From our first solvers on single-colored or colorless jigsaw puzzles, more methods aiming to solve the jigsaw puzzle based on its inside properties are in great desired.

7 Conclusion

In this project, we innovatively use feature maps from Resnet34 to assist one of the most important problems in jigsaw puzzle solving, adjacency prediction in the original image. Taking the low-level features and RGB information at edges, the classifier achieves comparable accuracy. Moreover, this can be used to solve single-colored or colorless jigsaw puzzles for the first time. The proposed algorithms can be used as the adjacency measurement in the end-to-end jigsaw puzzle solvers.

References

- [1] Tom Altman. Solving the jigsaw puzzle problem in linear time. *Applied Artificial Intelligence an International Journal*, 3(4):453–462, 1989.
- [2] Erik D Demaine and Martin L Demaine. Jigsaw puzzles, edge matching, and polyomino packing: Connections and complexity. *Graphs and Combinatorics*, 23(1):195–208, 2007.

- [3] Lucio Dery, Robel Mengistu, and Oluwasanya Awe. Neural combinatorial optimization for solving jigsaw puzzles: A step towards unsupervised pre-training. 2016.
- [4] Herbert Freeman and L Garder. Apictorial jigsaw puzzles: The computer solution of a problem in pattern recognition. *IEEE Transactions on Electronic Computers*, (2):118–127, 1964.
- [5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [7] David Koller and Marc Levoy. Computer-aided reconstruction and new matches in the forma urbis romae. *Bullettino Della Commissione Archeologica Comunale di Roma*, 2:103–125, 2006.
- [8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [9] Samuel R Levin. The computer and literary studies, 1975.
- [10] Marlos AO Marques and Cinthia OA Freitas. Reconstructing strip-shredded documents using color as feature matching. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 893–894. ACM, 2009.
- [11] Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.
- [12] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.
- [13] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 512–519. IEEE, 2014.
- [14] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint*, 2017.
- [15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
- [16] Dror Sholomon, Omid David, and Nathan S Netanyahu. A genetic algorithm-based solver for very large jigsaw puzzles. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1767–1774. IEEE, 2013.
- [17] Dror Sholomon, Omid E David, and Nathan S Netanyahu. Dnn-buddies: a deep neural network-based estimation metric for the jigsaw puzzle problem. In *International Conference on Artificial Neural Networks*, pages 170–178. Springer, 2016.
- [18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [19] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.