Convolutional Photomosaic Generation via Multi-Scale Perceptual Losses

Matthew Tesfaldet[1,2], Nariman Saftarli[3], Marcus A. Brubaker[1,2], and Konstantinos G. Derpanis[2,3]

[1] York University, Department of Electrical Engineering and Computer Science, Toronto, Canada
{mtesfald,mab}@eecs.yorku.ca

[2] Vector Institute, Toronto, Canada

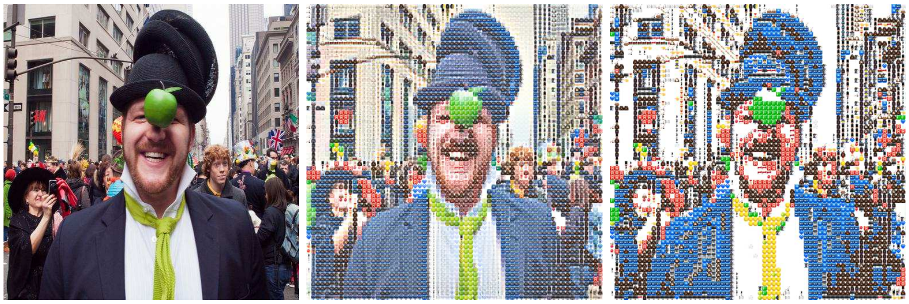
[3] Ryerson University, Department of Computer Science, Toronto, Canada
{nsaftarli,kosta}@scs.ryerson.ca

Abstract. Photographic mosaics (or simply *photomosaics*) are images comprised of smaller, equally-sized image tiles such that when viewed from a distance, the tiled images of the mosaic collectively resemble a perceptually plausible image. In this paper, we consider the challenge of automatically generating a photomosaic from an input image. Although computer-generated photomosaicking has existed for quite some time, none have considered simultaneously exploiting colour/grayscale intensity and the structure of the input across scales, as well as image semantics. We propose a convolutional network for generating photomosaics guided by a multi-scale perceptual loss to capture colour, structure, and semantics across multiple scales. We demonstrate the effectiveness of our multi-scale perceptual loss by experimenting with producing extremely high resolution photomosaics and through the inclusion of ablation experiments that compare with a single-scale variant of the perceptual loss. We show that, overall, our approach produces visually pleasing results, providing a substantial improvement over common baselines.

Keywords: Photomosaic, ASCII text, deep learning, perceptual loss, multi-scale analysis

1 Introduction

Photographic mosaics (or simply *photomosaics*) are images comprised of smaller, equally-sized image tiles (or "templates") such that when viewed from a distance, the tiled images of the mosaic collectively resemble a perceptually plausible image. Although the term has existed since the 1990s (specifically for photography), the unique art form of stitching together a series of adjacent pictures to produce a scene has existed since the 1970s. They are inspired from traditional mosaics, an ancient art form dating back at least as far as 1500 BCE, where scenes and



(a) Input

(b) Soft selection (during training)

(c) Hard selection (during inference)

Fig. 1: Given an input image, (a), and a collection of template images (pictured are 8×8 Apple emoji templates), our convolutional network generates a photomosaic, (c), that is perceptually similar to the input. For training our model, we exploit a continuous relaxation of the non-differentiable discrete template selection process to encourage the "soft" outputs, (b), to be as one-hot as possible for proper evaluation by our multi-scale perceptual metric. Zoom in for details.

patterns were depicted using coloured pieces of glass, stone or other materials. Here we focus on *computer-generated* photomosaics. Computer-generated photomosaicking relies on various algorithms to select suitable combinations of templates from a given collection to compose a photomosaic that is perceptually similar to a target image.

In early work, Harmon and Knowlton experimented with creating large prints from collections of small symbols or images. In their famous artwork, "Studies in Perception I" [6], they created an image of a choreographer by scanning a photograph with a camera and converting the grayscale values into typographic symbols. This piece was exhibited at one of the earliest computer art exhibitions, "The Machine as Seen at the End of the Mechanical Age", held at the Museum of Modern Art in New York City in 1968. Soon after, Harmon [7] investigated how much information is required for recognizing and discriminating faces and what information is the most important for perception. To demonstrate that very little detail was required for humans to recognize a face, he included a mosaic rendering of Abraham Lincoln consisting of varying shades of gray. Based on Harmon's findings, Salvador Dalí, in 1976, created the popular photomosaic, "Gala Contemplating the Mediterranean Sea" [4]. This was among the first examples of photomosaicking, and one of the first by a recognized artist.

Generally, there are two methods of photomosaicking: *patch-wise* (*e.g.*, [14]) and *pixel-wise* (*e.g.*, [18]). Patch-wise photomosaicking involves matching each tiled region with a template consisting of the closest average colour. In pixel-wise photomosaicking the matching is done on a per-pixel level between the pixels of the target image and the templates. This is computationally more expensive but generally produces more visually pleasing results since the per-pixel matching allows a rudimentary matching of structure.

Computer-generated photomosaicking has mostly been explored in the context of matching colour/grayscale intensities and, in an extremely limited sense, structures. Pixel-wise methods are limited to matching the colour of individual pixels, while patch-wise methods typically use simple similarity metrics that may miss important structural information, *e.g.*, edges, curves, etc. Both are limited to analysis at a single scale and generally ignore overall image semantics when producing a photomosaic. In contrast, our proposed approach involves a holistic analysis of colour, structure, and semantics across multiple scales.

Jetchev *et al.* [8] experimented with using convolutional networks (ConvNets) to form a perceptually-based mosaicking model; however, their approach was limited to a texture transfer process and consequently was not true photomosaicking, *i.e.*, their outputs did not consist of tiled images. Furthermore, their approach did not account for matching colours between the input and output, only structure, and only at a single scale.

In this paper, we propose a perceptually-based approach to generating photomosaics from images using a ConvNet. We rely on a perceptual loss [9] for guiding the discrete selection process of templates to generate a photomosaic. Inspired by previous work [17], we extend the perceptual loss over multiple scales. Our approach is summarized in Fig. 1.

We make the following contributions. Given a discrete set of template images, we propose a feed-forward ConvNet for generating photomosaics. To the authors' knowledge, we are the first to demonstrate a ConvNet for photomosaicking that utilizes a perceptual metric. We demonstrate the effectiveness of our multi-scale perceptual loss by experimenting with producing extremely high resolution photomosaics and through the inclusion of ablation experiments that compare with a single-scale variant of the perceptual loss. We show that, overall, our approach produces visually pleasing results with a wide variety of templates, providing a substantial improvement over common baselines.

2 Technical approach

Given an RGB input image, $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$, our goal is to generate a photomosaic, $\mathbf{Y} \in \mathbb{R}^{H \times W \times 3}$, where H and W denote the image height and width. For every non-overlapping tiled region in the image, we learn a distribution of weightings (or coefficients) for selecting templates. This is represented using a map of one-hot encodings, denoted by $\mathbf{C} \in [0,1]^{(H/H_T) \times (W/W_T) \times N_T}$, where H_T, W_T, and N_T denote the template height, template width, and the number of templates, respectively. Each spatial position on this map contains a one-hot encoding denoted by $\mathbf{c}_{r,c}$, where r and c correspond to its row and column position on the map. RGB templates, $\mathbf{T} \in \mathbb{R}^{H_T \times W_T \times 3N_T}$, are given and fixed between training and testing. In Section 2.1, we outline our encoder-decoder ConvNet architecture. Section 2.2 describes how we exploit a continuous relaxation of the argmax function to make training differentiable. Finally, Section 2.3 describes our multi-scale perceptual loss which is used to train the decoder portion of the function.

2.1 Encoder-decoder architecture

Our ConvNet is designed as an encoder-decoder network that takes $\mathbf{X}$ as input and produces $\mathbf{Y}$ as the photomosaic output. We adopt the VGG-16 [16] ConvNet pre-trained on the ImageNet dataset [15] as the encoder portion of our network, which is kept fixed. For the purpose of photomosaicking, we find using the layers up to `pool3` of VGG-16 to be sufficient. Our decoder is as follows: a $1 \times 1 \times 256$ (corresponding to $\textit{height} \times \textit{width} \times \textit{num\_filters}$) convolution, a ReLU activation, a $3 \times 3 \times N_T$ convolution (3×3 to encourage template consistency among neighbours), and a channel-wise softmax to produce the template coefficients. To keep the range of activations stable, we use layer normalization [2] after each convolution in the decoder. In all convolutional layers we use a stride of 1.

For each tiled region, $\mathbf{y}_{r,c}$, of the final output, $\mathbf{Y}$, let $\mathbf{c}_{r,c}(i)$ be the i-th coefficient of the one-hot encoding corresponding to that region and $\mathbf{T}(i) \in \mathbb{R}^{H_T \times W_T \times 3}$ the i-th template of RGB templates $\mathbf{T}$. The output $\mathbf{y}_{r,c}$ is generated by linearly combining the templates for that region by their respective template coefficients,

$$\mathbf{y}_{r,c} = \sum_{i=1}^{N_T} \mathbf{c}_{r,c}(i)\mathbf{T}(i) \ . \tag{1}$$

The final output, $\mathbf{Y}$, is a composition of each tiled output $\mathbf{y}_{r,c}$.

2.2 Learning a discrete selection of templates

Key to our approach is the *discrete* selection of templates at each tiled region. This is necessary to produce a photomosaic. During training, however, using an argmax to select the template with the maximal coefficient is not possible because the argmax function is non-differentiable. Instead, we exploit a continuous relaxation of the argmax by annealing the softmax that produces the coefficients. In particular, we gradually upscale the softmax inputs during training by $1/\tau$, where τ is the "temperature" parameter that is gradually "cooled" (*i.e.*, reduced) as training progresses. In the limit as $\tau \rightarrow 0$, the softmax function approaches the argmax function and Eq. 1 becomes nearly equivalent to a discrete sampler, as desired. Specifically, the softmax distribution of coefficients nears a one-hot distribution. This encourages the network to select a single template for each tiled region. During inference, however, instead of linearly combining templates by their respective coefficients, each tiled region output, $\mathbf{y}_{r,c}$, can be generated by selecting the template corresponding to the argmax of the distribution of coefficients, $\mathbf{c}_{r,c}$.

2.3 Multi-scale perceptual loss

So-called "perceptual losses" have previously been used as a representation of salient image content for image stylization tasks, *e.g.*, image style transfer [9,5].

Instead of generating images based on differences between raw colour pixel values, perceptual losses are used to enable high quality generation of images based on differences between low-level to high-level image feature representations extracted from the convolutional layers of a pre-trained ConvNet. To that end, we use a perceptual loss [9] to guide the network to produce photomosaics that are perceptually similar to the input. Specifically, the perceptual loss measures the difference between low-level features (*e.g.*, visual content such as edges, colours, curves) to high-level features (*e.g.*, semantic content such as faces and objects) computed on the input image and the output photomosaic. Like our encoder, we use the VGG-16 [16] ConvNet pre-trained on the ImageNet dataset [15]. However, here it is used as a perceptual metric and layers `conv1_1`, `conv2_1`, `conv3_1`, `conv4_1`, and `conv5_1` are used for computing the perceptual loss. Formally, let $\phi_l(\mathbf{X})$ be the activations of the l-th layer of VGG-16 when processing input $\mathbf{X}$. The perceptual loss is computed as the average Mean Squared Error (MSE) between feature representations of $\mathbf{X}$ and $\mathbf{Y}$,

$$L(\mathbf{X}, \mathbf{Y}) = \frac{1}{L} \sum_{l} ||\phi_l(\mathbf{X}) - \phi_l(\mathbf{Y})||_2^2 \ , \tag{2}$$

where L is the number of layers used for computing the perceptual loss.

To produce visually accurate photomosaics, we require the objective to consider the content within each tiled region as well as the content spanning multiple tiled regions. This necessitates analysis across multiple scales. Motivated by prior work [17,10], we compute the perceptual loss (Eq. 2) on a Gaussian pyramid [3] of the input and output. This guides the decoder to select templates that closely match the content within each tiled region, as well as collectively match the overall content of the input. To mitigate the influence of seams between tiled regions, we blur the photomosaic output before feeding it into the loss. Our final objective is as follows:

$$L(\mathbf{X}, B(\mathbf{Y})) = \frac{1}{SL} \sum_{s} \sum_{l} ||\phi_l(\mathbf{X}^s) - \phi_l(B(\mathbf{Y}^s))||_2^2 \ , \tag{3}$$

where input $\mathbf{X}^s$ is taken from the s-th level of a Gaussian pyramid, $B(\mathbf{Y}^s)$ is the blurred photomosaic output taken from the same level, and S is the number of scales used for the pyramid.

Training For training the weights of our decoder, we use the images from the Microsoft COCO dataset [13]. We train on a merger of the train, test, and validation splits of COCO. We resize each image to 512 × 512 and train with a batch size of 12 for 2,000 iterations. We use the Adam optimizer [11] with a learning rate of 6e$-$3 that is exponentially decayed every 100 iterations at a rate of 0.96. We follow a temperature cooling schedule starting from $\tau = 1$ and gradually decreasing τ every 10 iterations until $\tau = 0.067$. Our network is implemented using TensorFlow [1]. Training roughly takes 20 minutes on an NVIDIA Titan V GPU. Figure 2 shows results using various 8×8 templates on a 512×512 input.

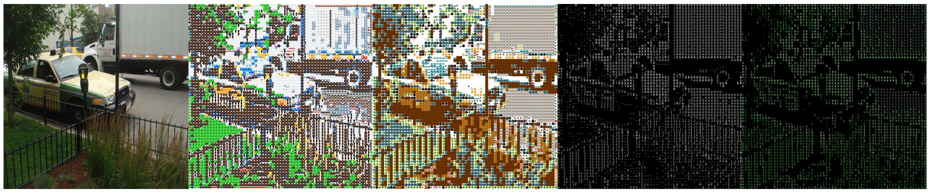


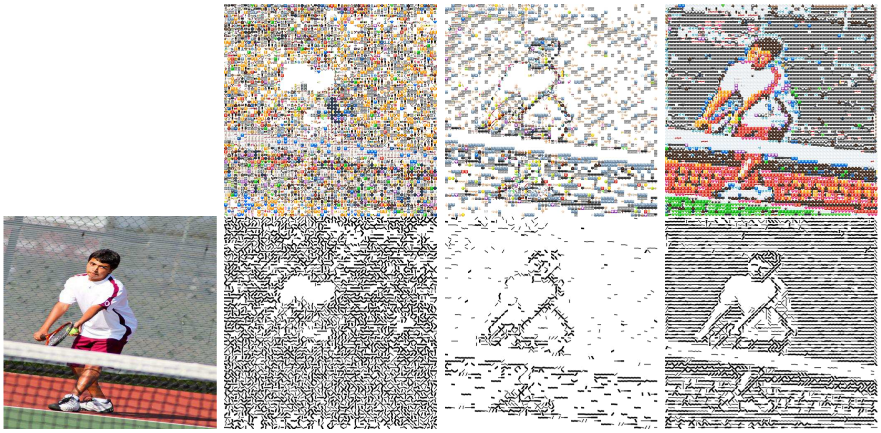
Fig. 2: Photomosaic results using 8 × 8 "glyphs" as templates. (left-to-right) Input, Apple emoji icons, sprites from "Super Mario Bros.", ASCII characters, text characters from "The Matrix". Zoom in for details.

3 Experiments

To evaluate our approach, we perform two experiments: a baseline qualitative comparison using nearest neighbour with both a simple L2 metric and with a Structural SIMilarity (SSIM) [19] metric, which is a perception-based metric that attempts to address shortcomings of L2 by taking the local image structure into account; and a qualitative comparison between using a single scale and multiple scales for the perceptual loss. Finally, we experiment with producing extremely high resolution photomosaics. For our full photomosaic results, collection of templates used, and source code, please refer to the supplemental material on the project website: ryersonvisionlab.github.io/perceptual-photomosaic-projpage.

Baselines To demonstrate that our approach improves upon common baselines in capturing colour, structure, and semantics across multiple scales, we compare against nearest neighbour with L2 and SSIM for template selection on two sets of templates: the complete set of emojis from Apple, and a specially-designed set of templates of oriented edges at varying thicknesses and rotations. Photomosaics are generated as follows: for each tiled region, the template with the lowest L2 loss or highest SSIM when compared with the underlying image content (in raw colour pixel values) is selected. Figure 3 shows our results. Nearest neighbour with L2 (Fig. 3b) completely fails in retaining both the colour and structure of the input. With SSIM (Fig. 3c), some structure of the input is preserved, albeit only at small scales, while colour accuracy is generally lacking. Moreover, both methods do not preserve the semantics of the input, such as the subject's hair, nose, and eyes. In contrast, our approach (Fig. 3d) reliably captures the colour, structure, and semantics of the image.

Single vs. multi-scale We perform an ablation study on our multi-scale perceptual loss to present the individual contributions of each scale (*i.e.*, fine and coarse) and to motivate the benefit of incorporating information across multiple scales. When the perceptual loss is operating on a single scale, it is restricted to scrutinizing the photomosaic output at that scale. As shown in Fig. 4, when the scale is only at a fine level, the output fails to preserve larger structures like the



(a) Input (b) NN with L2 (c) NN with SSIM (d) Ours

Fig. 3: Baseline comparisons. Given an input image, (a), photomosaics are generated using nearest neighbour (NN) with an L2 metric, (b), NN with a SSIM metric, (c), and our convolutional approach, (d). From (b) to (d), the top row of photomosaics consist of Apple emoji templates and the bottom row of photomosaics consist of oriented edge templates. Zoom in for details.

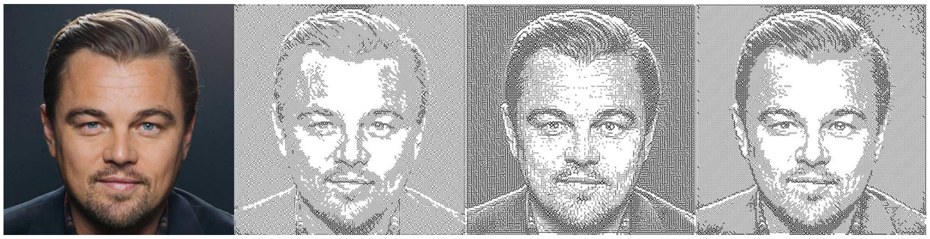


Fig. 4: Photomosaic outputs when using a single vs multi-scale perceptual loss. (left-to-right) Input, single-scale at a fine scale, single-scale at a coarse scale, multi-scale at both fine and coarse scales. Zoom in for details.

outline around the subject's jawline and ears. At a coarse level, the reduction in resolution prevents finer details from being captured, such as the orientation of edges in the input image, resulting in a noisier output. However, when using the multi-scale perceptual loss operating on both fine and coarse scales, the output reliably preserves both the finer details and the coarse structure of the image.

High resolution To demonstrate the effectiveness of using a multi-scale perceptual loss, we experiment with generating extremely high resolution photomosaics, as shown in Fig. 5. The input is a $5,280 \times 3,960$ image of Vincent Van Gogh's painting, "Starry Night", and the output is a visually compelling $10,560 \times 7,936$

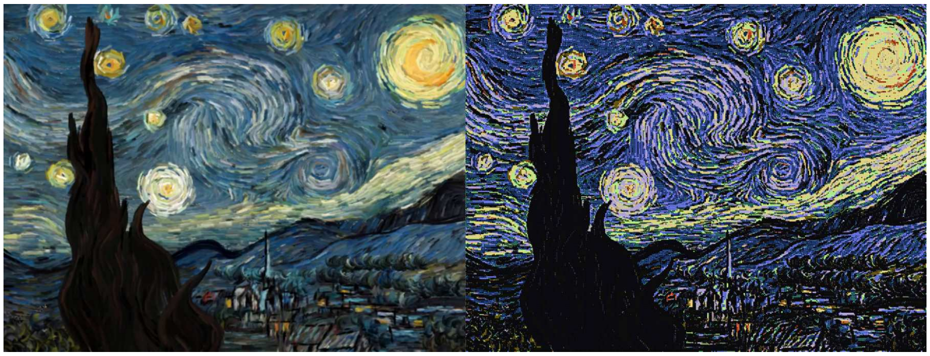


Fig. 5: High resolution photomosaics. (left) A $5,280 \times 3,960$ input and (right) a $10,560 \times 7,936$ photomosaic using 32×32 templates from a collection of 17,500 rotated and colour-shifted images taken from the top-100 images from the Hubble Space Telescope [12]. Shown are the downsampled versions of the images to save space; please see the supplemental for the full resolution images.

photomosaic. The multi-scale perceptual loss enables the model to capture both the coarse scale and fine scale features of the input. For example, the input image content spanning multiple tiled regions (*e.g.*, the large black tower and the stars) are reliably captured in the photomosaic through the appropriate composition of templates, while the input image content within tiled regions are reliably captured through the appropriate selection of templates that match the underlying image structure, such as the orientation and colour of the brush strokes.

4 Conclusion

In this paper, we presented a ConvNet for generating photomosaics of images given a collection of template images. We rely on a multi-scale perceptual loss to guide the discrete selection process of templates to generate photomosaics that best preserve colour, structure, and semantics of the input across multiple scales. We show that our approach produces visually pleasing results with a wide variety of templates, providing a substantial improvement over common baselines. We demonstrate the benefits of a multi-scale perceptual loss through the inclusion of ablation experiments and by experimenting with generating extremely high resolution photomosaics.

Acknowledgements MT is supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Canadian Graduate Scholarship. KGD and MAB are supported by NSERC Discovery Grants. This research was undertaken as part of the Vision: Science to Applications program, thanks in part to funding from the Canada First Research Excellence Fund.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: A system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16). pp. 265–283 (2016), https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf 5
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv:1607.06450 (2016) 4
3. Burt, P.J., Adelson, E.H.: The Laplacian pyramid as a compact image code. In: IEEE Trans. Commun. pp. 532–540 (1983) 5
4. Dalí, S.: Gala Contemplating the Mediterranean Sea which at Twenty Meters Becomes the Portrait of Abraham Lincoln Exhibited in 1976, Guggenheim Museum, New York 2
5. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR. pp. 2414–2423 (2016) 4
6. Harmon, L., Knowlton, K., Hay, D.: Studies in Perception I Exhibited at The Machine as Seen at the End of the Mechanical Age, Nov. 27, 1968 - Feb. 9, 1969, The Museum of Modern Art, New York 2
7. Harmon, L.D.: The recognition of faces. Scientific American **229**(5), 70–83 (1973) 2
8. Jetchev, N., Bergmann, U., Seward, C.: GANosaic: Mosaic creation with generative texture manifolds. In: NIPS Workshop (2017) 3
9. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV. pp. 694–711 (2016) 3, 4, 5
10. Ke, T.W., Maire, M., Yu, S.X.: Multigrid neural architectures. In: CVPR. pp. 6665–6673 (2017) 5
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2014) 5
12. Kornmesser, M.: Top 100 images. https://www.spacetelescope.org/images/archive/top100 (2015), images by ESA/Hubble (M. Kornmesser) 8
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755 (2014) 5
14. Martins, D.: Photo-mosaic. https://github.com/danielfm/photo-mosaic (2014), accessed: 2018-07-15 2
15. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015) 4, 5
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015) 4, 5
17. Snelgrove, X.: High-resolution multi-scale neural texture synthesis. In: SIGGRAPH ASIA Technical Briefs (2017) 3, 5
18. Tran, N.: Generating photomosaics: An empirical study. In: SAC. pp. 105–109 (1999) 2
19. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Signal Process pp. 600–612 (2004) 6