LEARNING VECTOR QUANTIZED SHAPE CODE FOR AMODAL BLASTOMERE INSTANCE SEGMENTATION

Won-Dong Jang[1] *Donglai Wei*[4] *Xingxuan Zhang*[1] *Brian Leahy*[1,2] *Helen Yang*[3]
James Tompkin[5] *Dalit Ben-Yosef*[6] *Daniel Needleman*[1,2] *Hanspeter Pfister*[1]

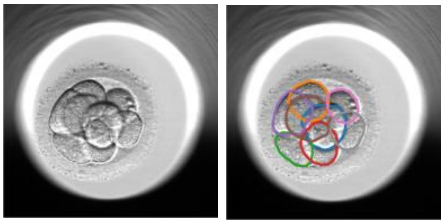
[1] School of Engineering and Applied Sciences, [2] Department of Molecular and Cellular Biology,
[3] Harvard Graduate Program in Biophysics, Harvard University, MA, USA
[4] Boston College, MA, USA
[5] Department of Computer Science, Brown University, RI, USA
[6] Tel-Aviv Sourasky Medical Center, Israel

ABSTRACT

Blastomere instance segmentation is important for analyzing embryos' abnormality. To measure the accurate shapes and sizes of blastomeres, their amodal segmentation is necessary. Amodal instance segmentation aims to recover an object's complete silhouette even when the object is not fully visible. For each detected object, previous methods directly regress the target mask from input features. However, images of an object under different amounts of occlusion should have the same amodal mask output, making it harder to train the regression model. To alleviate the problem, we propose to classify input features into intermediate shape codes and recover complete object shapes. First, we pre-train the Vector Quantized Variational Autoencoder (VQ-VAE) model to learn these discrete shape codes from ground truth amodal masks. Then, we incorporate the VQ-VAE model into the amodal instance segmentation pipeline with an additional refinement module. We also detect an occlusion map to integrate occlusion information with a backbone feature. As such, our network faithfully detects bounding boxes of amodal objects. On an internal embryo cell image benchmark, the proposed method outperforms previous state-of-the-art methods. To show generalizability, we show segmentation results on the public KINS natural image benchmark. Our method would enable accurate measurement of blastomeres in In Vitro Fertilization (IVF) clinics, potentially increasing the IVF success rate.

1. INTRODUCTION

Infertile couples worldwide use In-Vitro Fertilization (IVF) to treat their infertility. In a typical IVF treatment, clinicians visually inspect the embryos, select the one that appears most likely to form a viable pregnancy, and transfer it back to the mother. To aid in embryo selection, many modern clinics employ sophisticated time-lapse imaging systems [1]. One feature known to be predictive of an embryo's viability is the shape and symmetry among the cells in the early developing embryo, which are known as blastomeres [2]. However, current clinical practice visually scores the symmetry at a few distinct points in time, which is time-consuming, inaccurate, and omits much information about the embryo, especially when time-lapse imaging is used. This makes replacing visual symmetry scoring with automated blastomere segmentation (Fig. 1a) a prime candidate for improving clinical IVF practice.



(a) Blastomere Segmentation

(b) Previous Approach (c) Our Approach

Fig. 1. **Amodal Blastomere instance segmentation.** (a) We show an image and its amodal segmentation masks for translucent blastomere cells overlapping with each other. (b) Previous approaches directly regress the amodal mask from the region of interest (ROIAlign) features. (c) Instead, we first learn a vector quantized (VQ) shape code from ground truth amodal masks, and then classify ROIAlign features into these discrete codes.

However, while clinics have collected many embryo images from IVF cycles, most existing blastomere segmentation algorithms [3–7] use hand-crafted features instead of data-driven approaches. Since hand-crafted methods are tailored to a certain dataset, they may not be robust on different datasets collected in varying environments. Moreover, they do not take into account amodal visual reconstruction, predicting the complete shape of partially-visible objects.

Many recent deep-learning-based models have been proposed for amodal segmentation [8–13]. However, these approaches often do not have prior knowledge of the underlying shape, which makes the shape difficult to predict from instance observations under different amounts of occlusion. Further, unlike normal instance segmentation, images of an object under different amounts of occlusion

This work has been completed while Won-Dong Jang were in Harvard University. Xingxuan Zhang contributed to this work while he was an intern student in Harvard University.

should have the same amodal mask output. Thus, it will be more robust to classify input features into an intermediate robust representation instead of working on the pixel-level. Fig. 1a visualizes a pair of an embryo image and its ground truth. The cells highly overlap each other, but their underlying shapes are still predictable.

To exploit this additional information, we propose to learn discrete supervised learning amodal instance segmentation algorithm for partially-visible objects. From binary masks of our object class, we create a deep shape prior as an embedding space with a vector quantized-variational autoencoder (VQ-VAE; [14]). Then, we train our segmentation model to predict the latent representation of an object mask in a bounding box (Fig. 1c).

Segmentation performance of proposal-based instance segmentation methods [15, 16] highly depends on the bounding box quality. In amodal segmentation, occlusion makes having accurate bounding boxes even more difficult. To tackle this occlusion problem, we add an occlusion detection module to a backbone network. This allows our network to propose better bounding boxes by integrating the occlusion information with the backbone features.

We first experiment with a real embryo cell biomedical dataset. Then, we conduct experiments on natural images of street scenes via the KINS dataset [17] to show the generalizability of our method. Our approach of encoding objects outperforms state-of-the-art instance segmentation algorithms [13, 15] on both the translucent and occluded types of tested partial visibility. In summary, our contribution is to propose a novel formulation that incorporates a vector quantized shape code into the amodal instance segmentation pipeline. Additionally, we exploit occlusion information when detecting and segmenting amodal objects via occlusion detection, which can be a new direction for amodal segmentation. This method achieves state-of-the-art performance on not only an internal biomedical image dataset but also the KINS natural image dataset.

2. RELATED WORKS

Blastomere Segmentation: Traditional methods predict semantic blastomere masks using hand-crafted features without the instance-level segmentation. Khan *et al.* [7] set seeds inside and outside of cells and optimize Markov random field for segmentation. Rad *et al.* [3] and Kheradmand *et al.* [6] generate blastomere candidates from extracted edges and select the best candidate in terms of edge coverage. Sidhu and Mills [4] apply thresholding and morphological operations to find the regions of blastomeres and find centers of each cell by measuring distances from pixels to the closest boundary.

Cell-Net proposed by Rad *et al.* [5] is the closest method to ours, training a convolutional neural network for cell localization. Recently, five convolutional neural networks have been trained for key morphological feature extraction [18], including blastomere segmentation. However, Cell-Net only predicts blastomere centers while we perform amodal instance segmentation. Even though the key morphological feature extraction method outputs masks of blastomeres, it focuses on finding the best setting for the input data using existing Mask R-CNN [15].

Amodal Instance Segmentation: Li and Malik [8] introduce the first amodal segmentation method. They predict bounding-boxes of modal parts of objects using the object detector [19] and extract segmentation masks using a neural network accepting a pair of an image and a bounding-box as the input. Qi *et al.* [13] present an amodal segmentation dataset, KINS, by annotating the KITTI detection dataset. They also propose an amodal segmentation network by adding occlusion classification and amodal segmentation branches to the Mask R-CNN framework [15].

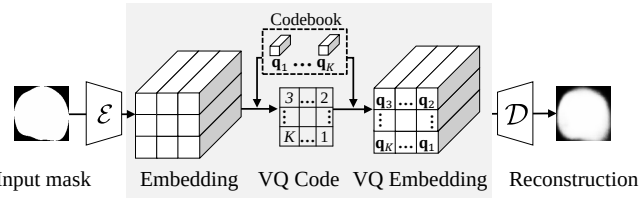


Fig. 2. VQ-VAE architecture containing the mask encoder, embedding quantizer, and mask decoder networks.

Vector Quantization in Deep Learning: Vector quantization methods have been widely used for image compression [20, 21]. van den Oord *et al.* [22] proposed a vector quantized variational autoencoder for image generation. They show that the proposed method generates more realistic images using learned template codewords.

3. VECTOR QUANTIZED SHAPE CODE

Our goal is to learn a discrete representation of amodal shape masks. With it, we can re-formulate the amodal instance segmentation as a classification problem in the low-dimensional latent space. Comparing to previous dense pixel-level mask prediction, the proposed approach can be robust to occlusion changes and regularized in geometry. To this end, we train a vector quantized variational autoencoder (VQ-VAE) model on the amodal masks to learn the vector quantized (VQ) shape code.

Comparing Latent Variable Models: To learn a compact representation of the input, variational autoencoder models (VAE) [14] are commonly used with the Gaussian prior distribution of the latent variable. VAEs learn a global continuous code of the input with the mask encoder model $\mathcal{E}$, which can be decoded back for input reconstruction with the mask decoder model $\mathcal{D}$. To discretize the learned code, VAE-based clustering methods jointly learn a codebook of embedding vectors that serve as clustering centers. However, as the learned embedding is global, it takes a large codebook for the input to find a similar quantized code. It requires an even larger codebook for a larger number of object categories. VQ-VAEs [22] predict embeddings with spatial resolution and jointly learn a global codebook (Fig. 2). With it, we can use the quantized embeddings to reconstruct input with a limited codebook size.

VQ-VAE Model: The key component of VQ-VAE models is the embedding quantizer module. During inference, the mask encoder first transforms the input binary mask $\mathbf{x}$ into a set of latent vectors $\mathbf{e}$. Then, the embedding quantizer assigns each latent vector to the nearest code in the pre-trained codebook $\{\mathbf{q}_1, \ldots, \mathbf{q}_K\}$. Lastly, the mask decoder transforms the quantized embeddings $\hat{\mathbf{e}}$ back into a binary mask.

Learning: The loss function combines a reconstruction loss, a codebook loss, and a commitment loss. The reconstruction loss is defined as the cross-entropy loss between input mask $\mathbf{x}$ and the reconstructed mask $\mathcal{D}(\hat{\mathbf{e}})$. The codebook loss, which only applies to the codebook, makes the selected codes $\hat{\mathbf{e}}$ close to the predicted latent vector $\mathbf{e}$. The commitment loss, which only applies to the mask encoder, forces the latent vectors $\mathcal{E}(\mathbf{x})$ to stay close to the matched codes to prevent excessive fluctuations of codes. The full VQ-VAE loss function $\mathcal{L}_{\text{v}}$ is

$$\mathcal{L}_{\text{v}} = \|\mathbf{x} - \mathcal{D}(\hat{\mathbf{e}})\|_2^2 + \|[\mathbf{e}] - \hat{\mathbf{e}}\|_2^2 + \beta\|\mathcal{E}(\mathbf{x}) - [\hat{\mathbf{e}}]\|_2^2, \quad (1)$$

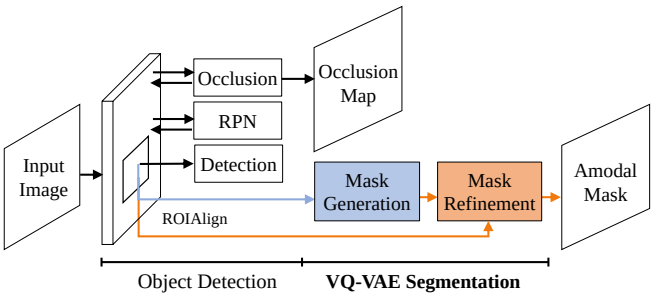


Fig. 3. **Overview of amodal segmentation pipeline.** We start from an instance segmentation pipeline, *e.g.*, Mask-RCNN. We add the occlusion detection module and replace the original FCN with the proposed VQ-VAE segmentation module. The proposed segmentation model has two steps: initial mask generation through VQ shape code prediction and mask refinement for better localization.

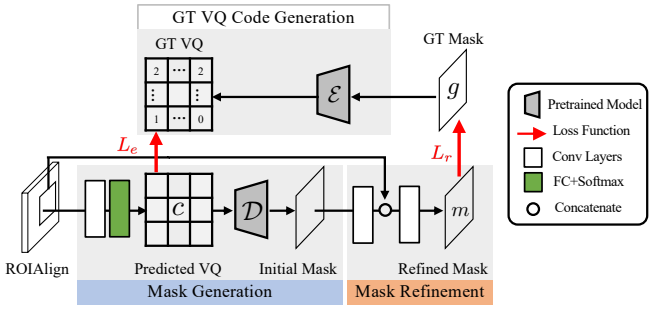


Fig. 4. **VQ-VAE segmentation module.** We have two segmentation stages: mask generation and mask refinement. We simultaneously minimize the two loss functions, $\mathcal{L}_{\text{e}}$ and $\mathcal{L}_{\text{r}}$.

where the operator [.] stands for a stop gradient operation that blocks gradients from flowing into its argument, and β is a hyper-parameter, which is set to 0.25.

4. AMODAL INSTANCE SEGMENTATION PIPELINE

We propose the VQ-VAE segmentation module to improve amodal instance segmentation. We take the proposal-based instance segmentation approach that contains two modules: object detection and mask prediction (Fig. 3). We attach an occlusion detection branch to object detection (Sec. 4.1) and replace previous fully convolutinal network (FCN) with the proposed module for mask prediction (Sec. 4.2). The whole pipeline is trained end-to-end (Sec. 4.3).

4.1. Object Detection Module

Unlike Faster-RCNN [19], our detection module predicts both bounding boxes and a binary occlusion map. Detecting locations of occlusions allows our detection module to predict accurate bounding boxes for partially visible objects. Using the backbone features, we estimate probabilities of each pixel being occluded $\{d_i\}$ via four convolution layers. We adopt the binary cross entropy loss:

$$\mathcal{L}_{\text{o}} = -\sum_{i \in H \times W} \{l_i \log d_i + (1 - l_i) \log (1 - d_i)\}, \quad (2)$$

where $H \times W$ is the spatial resolution of the backbone feature map and l_i is the ground-truth occlusion label at pixel i. We concatenate the output of the second-to-last convolution layer and the backbone feature map to exploit occlusion information in the detection and segmentation modules.

4.2. VQ-VAE Segmentation Module

As shown in Fig 4, the proposed VQ-VAE segmentation module has two steps: initial mask generation and mask refinement. It first generates an initial mask by decoding the predicted VQ-VAE shape code. Then, the refinement step learns to better align the initial mask with the visible object boundaries.

Initial Mask Generation: Given the instance-level feature from the object detection module, we first predict the vector quantized shape code and use a pre-trained VQ-VAE decoder model to decode it into object masks with complete shapes. We first predict a vector quantized shape code instead of a pixel-level binary mask to capture complete shapes using VQ-VAE. We use three convolution layers and one fully connected layer to predict codewords of vector quantized shape code $\mathbf{c}$. We formulate the problem of vector quantized shape code prediction as a classification problem. For the classification target, we use the pre-trained VQ-VAE mask encoder $\mathcal{E}$ to encode the ground truth instance mask $\mathbf{g}$ as shown in the right block in Fig 4. One hot encoding makes the encoded mask $\mathcal{E}(\mathbf{g})$ as a binary representation $\mathbf{b}$. For the codeword classification at each spatial location, the binary cross entropy loss is defined as

$$\mathcal{L}_{\text{e}} = -\sum_{i \in M \times M \times K} \{b_i \log c_i + (1 - b_i) \log (1 - c_i)\}, \quad (3)$$

where $M \times M$ is a spatial resolution of a vector quantized shape code and K is the number of codewords. We then feed the predicted VQ shape code $\mathbf{c}$ into the VQ-VAE mask decoder $\mathcal{D}$ to obtain an initial mask.

Mask Refinement: The vector quantized shape code can be powerful for shape completion, but the initial mask may not be well-aligned with the detailed object boundary.We add another mask refinement step that combines the instance-level feature and the initial mask feature. To train the refinement decoder, we set its loss function as

$$\mathcal{L}_{\text{r}} = -\sum_{i \in N \times N \times C} w_i \{g_i \log(m_i) + (1 - g_i) \log(1 - m_i)\}, \quad (4)$$

where m_i is the probability of a target object occurring at pixel i. $N \times N$ is a spatial resolution of the output mask and C indicates the number of object categories. The weight w_i is 1 for the channel of the ground-truth object class, otherwise 0.

4.3. Learning Strategy

During training, parameters in the region proposal network, detection, mask generation, and refinement modules are updated together to minimize the sum of the loss functions: $\mathcal{L} = \mathcal{L}_{\text{p}} + \alpha\mathcal{L}_{\text{d}} + \beta\mathcal{L}_{\text{o}} + \gamma\mathcal{L}_{\text{e}} + \delta\mathcal{L}_{\text{r}}$, where $\mathcal{L}_{\text{p}}$ and $\mathcal{L}_{\text{d}}$ indicate the losses for the region proposal network and the detection module, respectively. Hence, we train the proposed network in an end-to-end manner. Empirically, we set the hyper-parameters $\alpha = \gamma = \delta = 1$ and $\beta = 0.01$.

Table 1. Comparison of mAP on the embryo cell dataset.

Metric	FCN [15]	VQ-VAE (ours)
mAP	66.5%	**68.2%**

Table 2. Comparison of mAP metric on the KINS dataset [13].

Detection	FCN	VQ-VAE (ours)
Mask R-CNN	29.3% [15]	**30.3%**
Mask R-CNN + ASN	31.1% [13]	**31.5%**

5. EXPERIMENTS

We compare the proposed method with state-of-the-art methods on a microscopy image dataset and a natural image dataset. We then perform ablation studies on the natural dataset to better understand each component and validate our design choices.

5.1. Experiment Setup

Comparison methods: For amodal instance segmentation, we can use different object detection pipelines, *e.g.*, Mask R-CNN [15]. With the same pipeline, the proposed VQ-VAE segmentation module is compared with the fully convolutional network (FCN).

Metrics: We use mean average precision (mAP), which is standard for object instance segmentation [23]. Let AP_k denotes a predicted segmentation as correct if its mask intersection over union (IoU) is higher than k. mAP score is the average of $\{\text{AP}_k\}$ where k ranges from 0.5 to 0.95 at 0.05 intervals.

5.2. Main Results on Embryo Cell Images

IVF clinicians predict embryo transfer success by visually observing cell properties like size, granularity, and cleavage (cell split) timing. Cell segmentation of embryo images would automate this property collection for more efficient prediction. Note that our method is more interpretable by clinicians compared to predicting a single number (cell count) from the input image [24].

Data: We use subsets of embryo images [18] whose number of cells varies from two to eight. Each image is with a spatial resolution of 500×500 pixels. Note that we exclude one cell images to evaluate amodal instance segmentation methods. We use 7,054 images for training and 4,617 for testing. We find that cells are highly overlapping and only partially visible. The size of cells varies as cells cleave and shrink.

Results: Table 1 compares the results of our proposed algorithm with Mask R-CNN [15]. We report mean average precision (mAP) for the evaluation of the cell segmentation methods. The proposed algorithm outperforms the baseline methods by 1.7%. We believe the performance gain is from the vector quantized segmentation module and the occlusion detection. The vector quantized segmentation module makes predicted masks to be in-distribution so that we can always have blastomere-looking masks. On the other hand, learning occlusion detection flourishes the embeddings from the backbone.

5.3. Additional Results on Natural Images

To demonstrate our method's general applicability, we test on an amodal instance segmentation dataset for natural images with a greater diversity of shapes.

Table 3. Ablation study on the KINS dataset [13].

Setting	mAP
VQ-VAE	28.1%
VQ-VAE + Refinement	29.8%
VQ-VAE + Refinement + Occlusion map	**30.3%**

Data: The KINS dataset [13] is a benchmark for amodal instance segmentation algorithms, which is originally from the KITTI dataset [25]. It consists of 7,474 training and 7,517 test images of driving scenes. The annotated objects belong to one of 7 object classes: pedestrian, cyclist, car, van, tram, truck, and misc-vehicle. The KINS dataset provides both amodal and inmodal ground-truth annotations.

Results: Table 2 lists mean average precision metrics of the results of the proposed algorithm with Mask R-CNN [15] and Mask R-CNN + ASN [13]. Our proposed algorithm performs better than the conventional FCN method on both Mask R-CNN and Mask R-CNN + ASN pipelines. It demonstrates that our model is generalizable and thus can be applied to other amodal segmentation tasks.

Ablation Study: We perform two ablation studies on the KINS dataset. We chose KINS over the embryo dataset for more general analysis since the objects in KINS have more diverse shapes. We use Mask R-CNN in these studies. First, we remove the occlusion detection branch (VQ-VAE + Refinement). To this end, we train the network without the loss function for occlusion detection $\mathcal{L}_{\text{o}}$. Second, we exclude the refinement decoder in the segmentation module (VQ-VAE). To train the network without the refinement decoder, we minimized the embedding loss $\mathcal{L}_{\text{e}}$ only. We compare these two settings with the full architecture (VQ-VAE + Refinement + Occlusion map) on the KINS dataset. Table 3 lists the mAP scores for each ablation setting. Our full architecture performs 0.303 mAP, which is better than the other settings. It indicates that all our components are necessary for accurate amodal segmentation. The inferior performance of the setting without refinement comes from the lack of low-level features.

6. CONCLUSION

We proposed an image segmentation method for blastomere instances, which outputs complete masks of cells automatically. We show that it is effective to learn a mapping from the bounding box features to a shape prior embedding space from a VQ-VAE. This allows us to cope with translucent cells. We also show the benefits of occlusion detection for amodal object detection and segmentation. Our method is applicable for any partially visible objects, not only cells but also geometric shapes, cars, or pedestrians. Experimental results on the embryo and KINS demonstrated that our proposed algorithm outperforms state-of-the-art object instance segmentation methods [13, 15].

Our future works include application to other objects in natural scenes and expanding to biomedical problems that suffer occlusions, such as human blood cell segmentation. We also suggest proposal-free amodal segmentation networks with the center prediction to achieve real-time running speed. Lastly, by adopting generative adversarial networks [26], we might be able to learn shape priors better.

Acknowledgment

This work was funded in part by NIH grants 5U54CA225088 and R01HD104969, NSF Grant NCS-FO 1835231, the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard (award number 1764269), the Harvard Quantitative Biology Initiative, and Sagol fund for studying embryos and stem cells; Perelson Fund.

Compliance with Ethical Standards

This research study was conducted retrospectively using human subject data made available in restricted access by Tel-Aviv Medical Center, Israel. Ethical approval was not required as we reused previously introduced data.

7. REFERENCES

[1] Sarah Armstrong, Priya Bhide, Vanessa Jordan, Allan Pacey, Jane Marjoribanks, and Cindy Farquhar, "Time-lapse systems for embryo incubation and assessment in assisted reproduction," *Cochrane Database of Systematic Reviews*, , no. 5, 2019, Publisher: John Wiley & Sons, Ltd.

[2] Catherine Racowsky, Judy E. Stern, William E. Gibbons, Barry Behr, Kimball O. Pomeroy, and John D. Biggers, "National collection of embryo morphology data into Society for Assisted Reproductive Technology Clinic Outcomes Reporting System: associations among day 3 cell number, fragmentation and blastomere asymmetry, and live birth rate," *Fertility and Sterility*, vol. 95, no. 6, pp. 1985–1989, 2011, Publisher: Elsevier.

[3] Reza Moradi Rad, Parvaneh Saeedi, Jason Au, and Jon Havelock, "A hybrid approach for multiple blastomeres identification in early human embryo images," *Computers in biology and medicine*, vol. 101, pp. 100–111, 2018, Publisher: Elsevier.

[4] Simarjot S. Sidhu and James K. Mills, "Automated Blastomere Segmentation for Early-Stage Embryo Using 3D Imaging Techniques," in *ICMA*, Aug. 2019.

[5] Reza Moradi Rad, Parvaneh Saeedi, Jason Au, and Jon Havelock, "Cell-Net: Embryonic Cell Counting and Centroid Localization via Residual Incremental Atrous Pyramid and Progressive Upsampling Convolution," *IEEE Access*, vol. 7, pp. 81945–81955, 2019.

[6] Shakiba Kheradmand, Parvaneh Saeedi, Jason Au, and John Havelock, "Preimplantation Blastomere Boundary Identification in HMC Microscopic Images of Early Stage Human Embryos," in *arXiv:1910.05972 [cs, eess, q-bio]*, Oct. 2019, arXiv: 1910.05972.

[7] Aisha Khan, Stephen Gould, and Mathieu Salzmann, "Segmentation of developing human embryo in time-lapse microscopy," in *ISBI*, Apr. 2016.

[8] Ke Li and Jitendra Malik, "Amodal instance segmentation," in *ECCV*, 2016.

[9] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár, "Semantic amodal segmentation," in *CVPR*, 2017.

[10] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi, "SeGAN: Segmenting and generating the invisible," in *CVPR*, 2018.

[11] Patrick Follmann, Rebecca Kö Nig, Philipp Hä Rtinger, Michael Klostermann, and Tobias Bö Ttger, "Learning to see the invisible: End-to-end trainable amodal instance segmentation," in *WACV*, 2019.

[12] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G Schwing, "SAIL-VOS: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines," in *CVPR*, 2019, pp. 3105–3115.

[13] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia, "Amodal instance segmentation with KINS dataset," in *CVPR*, 2019, pp. 3014–3023.

[14] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," in *ICLR*, 2013.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2961–2969.

[16] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia, "Path aggregation network for instance segmentation," in *CVPR*, 2018, pp. 8759–8768.

[17] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 961–972, 2018, Publisher: Springer.

[18] Brian D. Leahy, Won-Dong Jang, Helen Y. Yang, Robbert Struyven, Donglai Wei, Zhe Sun, Kylie R. Lee, Charlotte Royston, Liz Cam, Yael Kalma, Foad Azem, Dalit Ben-Yosef, Hanspeter Pfister, and Daniel Needleman, "Automated Measurements of Key Morphological Features of Human Embryos for IVF," in *MICCAI*, 2020.

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015, pp. 91–99.

[20] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc V Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations," in *Advances in neural information processing systems*, 2017, pp. 1141–1151.

[21] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár, "Lossy image compression with compressive autoencoders," *arXiv preprint arXiv:1703.00395*, 2017.

[22] Aaron van den Oord, Oriol Vinyals, and others, "Neural discrete representation learning," in *Advances in neural information processing systems*, 2017, pp. 6306–6315.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014, pp. 740–755, tex.organization: Springer.

[24] Aisha Khan, Stephen Gould, and Mathieu Salzmann, "Deep convolutional neural networks for human embryonic cell counting," in *ECCV Workshops*, 2016.

[25] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *CVPR*, 2012.

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NeurIPS*, 2014, pp. 2672–2680.