## **Paper Decision**

*CVPR 2023 Conference Program Chairs*

28 Feb 2023CVPR 2023 Conference Paper3238 DecisionReaders: Program Chairs, Paper3238 Senior Area Chairs, Paper3238 Area Chairs, Paper3238 Reviewers Submitted, Paper3238 Authors

**Decision:**Reject

**Comment:**The reviewers expressed the following concerns prior to the rebuttal: \* Issues with the writing and visual results that impact clarity and comprehension (3hfa, 9Nvx, c46M) \* Unfair comparisons, e.g. training from scratch vs pre-train (3hfa, 9Nvx, c46M) \* Lack of discussion of computational cost (9Nvx, c46M) \* Incorrect claims (c46M) Unfortunately, the rebuttal raised several additional questions as reviewers noted that it did not sufficiently address their concerns and caused confusion: \* Difficulty interpreting the new results (3hfa) \* Lack of clarity around what changes will be made and what impact that will have on the paper (3hfa) \* SQL's large increase in parameters (9Nvx) \* Impact of backbones (c46M) The reviewers engaged in extensive discussion attempting to understand the new results and the inconsistencies. They indicated that they read the rebuttal when making their final decisions. While they were initially positive, in the end, there was limited support for the paper, with all reviewers voting for weak reject. This AC agrees with the weaknesses outlined by the reviewers as there are multiple questions that need to be clarified to ensure that the comparisons are fair and that the performance gains can be attributed to the main contributions of the paper. As a result, the final recommendation of this AC is reject. The authors are encouraged to take the constructive comments of the reviewers onboard in any future revision.

[[–]](https://openreview.net/forum?id=fF2UW7W4dR5" \o "Collapse reply thread)

*CVPR 2023 Conference Paper3238 Authors[Youhong Wang](https://openreview.net/profile?id=~Youhong_Wang1)(privately revealed to you)*

29 Jan 2023 (modified: 01 Feb 2023)CVPR 2023 Conference Paper3238 RebuttalReaders: Program Chairs, Paper3238 Senior Area Chairs, Paper3238 Area Chairs, Paper3238 Reviewers Submitted, Paper3238 Authors

[[–]](https://openreview.net/forum?id=fF2UW7W4dR5" \o "Collapse reply thread)

## **Official Review of Paper3238 by Reviewer 3hfa**

*CVPR 2023 Conference Paper3238 Reviewer 3hfa*

16 Jan 2023 (modified: 08 Feb 2023)CVPR 2023 Conference Paper3238 Official ReviewReaders: Program Chairs, Paper3238 Senior Area Chairs, Paper3238 Area Chairs, Paper3238 Reviewers Submitted, Paper3238 Authors

**Paper Summary:**

The paper proposes novel network architecture for the task of monocular depth estimation. The method trains a depth and pose networks using standard combination of L1 and SSIM losses using monocular video sequences. The depth network first uses encoder-decoder network to predict features for every pixel position. The features are further convolved with a large kernel and stride to estimate patch-level features. The patch-level features are pass through transformer layers to form "queries" that are used to build self-cost volume (a better name would be a 2D feature volume) using dot-product between "queries" and pixel-level features. The self-cost volume is further used to regress depth bin boundaries for the whole image and to predict depth bin probabilities for each pixel. The final depth is a weighted average of depth bins weighted by depth bin probabilities. Experiments on KITTI and Cityscapes show great performance of the model and qualitative results on Make3D demonstrate generalization ability. Ablation study demonstrates that the proposed network modules improve scores.

**Paper Strengths:**

* The reported results on KITTI benchmark are state of the art. Cityscape and Make3D scores are very good too.
* There are evaluations on 3 datasets and ablation study explores different design choices.
* The related work is comprehensive and results section has many of recent works.
* The qualitative results (many more in the supplementary materials) show very sharp and high-quality depth maps.

**Paper Weaknesses:**

* The writing has many small issues regarding grammar, word choices and details. Unfortunately, this also means that important details of the architecture are not completely clear. For example, "self-cost volume" is not a good name for a 2D feature map ("volume" typically refers to an array with 3 spatial dimensions like X, Y, Z). "Queries" Q in L364 seem like queries in a transformer architecture. Which means that they should be multiplied with "keys" and pass through softmax layer to produce "attention". However, in equation (3) there are no keys or softmax layer. The chosen terminology confuses the reader instead of helping. See more details in "additional comments for authors" section.
* The evaluation on Cityscapes is not completely fair. In L633-645 the authors use a KITTI-trained model and fine-tune on Cityscapes, while some of competition trains from scratch. The authors claim that this shows improved generalization of their method, which is not a fair claim as additional data even when pre-training have beneficial impact on many tasks (see transfer learning). This should be pointed out in the evaluation tables.
* The models are trained on monocular video sequences, so they lack metric scale information. There were no details about scale estimation during testing.
* Many methods in Table 1 try to have the same number of parameters in the networks for fair comparison. For example, Monodepth2 has results for Resnet18 and Resnet50 models. The lack of details on the network architecture and number of parameters make it difficult to compare to previous work.

**Overall Recommendation:**3: borderline

**Justification For Recommendation And Suggestions For Rebuttal:**

The performance on KITTI is great, however the current state of the text and lack of details are a problem.

**Confidence Level:**3: The reviewer is fairly confident that the evaluation is correct.

**Additional Comments For Authors:**

Here are suggestions to authors regarding text:

* L010 there is no paper ID
* L021 "the failure of" -> "failures in"
* L023 "namely" -> "named"
* L024 "make efficient use" -> "make use"
* L028 "the depth map in a probabilistic combination manner using" -> "the depth map probabilistically using"
* L029 "attention approximation" -> "features".
* L035 "generalization ability against" -> "generalization and robustness to"
* L050 & L052: You list shortcomings when you say "One the one hand" and "One the other hand", which is incorrect. Consider using "It is time consuming... world. Additionally, supervised methods..."
* L099 "generalization ability" -> "generalization"
* L101 "approachs" -> "approaches"
* L102 "unlabled" -> "unlabeled"
* L104 "existing methods are focusing" -> "existing methods have focused". Please choose either past tense or present tense.
* L108 "Besides" -> "Furthermore"
* L111 "make efficient use" -> "make use"
* L122 "depth value" -> "depths"
* L122 "gradually changes" -> "change gradually"
* L127 "To specify" -> "Specifically"
* L132 "innovative" -> "novel"
* L135 "innovative" -> "novel"
* L136 "volumn" -> "feature map"
* L145 "distillated" -> "distilled"
* L145 "used auxiliary" -> "using aurxiliary"
* L149 "generalization ability" -> "generalization"
* L153 "showed" -> "shows"
* L154 "generality" -> "generalization"
* L161 "a RGB image" -> "an RGB image"
* L163 "fomulate" -> "formulate"
* L178 "temporary" -> "temporal"
* L211 "temporary" -> "temporal"
* L240 "Methodology" -> "Method"
* L245 "firstly" -> "first"
* L256 "firstly" -> "first"
* L298 "ellaborated" -> "elaborate"
* L323 "Idealy" -> "Ideally"
* L358 "To specify" -> "Specifically"
* L381 "To specify" -> "Specifically"
* L704 "contains has" -> "contains"
* L749 "generality" -> "generalization"
* L780 "embeddings of 4x4 patches" -> "embeddings of 4x4 sized patches"
* L781 "embeddings of 16x16 patches" -> "embeddings of 16x16 sized patches"
* L792 "reception field" -> "receptive field"

Many of these issues can be fixed with a spellchecker or similar tool.

Many references do not have correct format. For example:

* no capitalization in names of conferences
* no capitalization in some paper names
* arxiv and conference both present as venue
* the same conference has many names,
* [15] and [16] are the same paper
* the formatting of conferences is inconsistent: some have pages, some have volume, some have editors, etc.

The colormap used in visualization of depth maps is different from prior work. It is not just "plasma", there is some resampling of colors. See monodepth2 depth map visualization function.

**Final Rating:**2: weak reject

**Final Rating Justification:**

I thank authors for the answers in the rebuttal. Unfortunately, I have to recommend "reject" at this point.

First, it was difficult to interpret the new results mentioned in the rebuttal (is there MMask for green row? why is there no MMask for Monodepth2? Why is yellow row performing so poorly? What happens if you use Resnet18 and Resnet50 as backbone?).

Furthermore, there are too many small issues with the paper and the rebuttal that it is difficult to judge what the final paper would look like. For example, text of A2 in rebuttal is confusing, I don't follow the argument and the lack of details of baselines makes me think that the text in the final paper would still be confusing. Another example is in the A3. The text suggests that the figure will make the architecture more clear, but I'm still confused about the choice of Q (introduced in paper L364, just a few lines below reference to Vision Transformer in L356. Text in A3 says there is no attention, yet Eq 4 is doing attention based aggregation). I'm just too uncertain about the details to confidently recommend the paper for acceptance.

[[–]](https://openreview.net/forum?id=fF2UW7W4dR5" \o "Collapse reply thread)

## **Official Review of Paper3238 by Reviewer 9Nvx**

*CVPR 2023 Conference Paper3238 Reviewer 9Nvx*

11 Jan 2023 (modified: 08 Feb 2023)CVPR 2023 Conference Paper3238 Official ReviewReaders: Program Chairs, Paper3238 Senior Area Chairs, Paper3238 Area Chairs, Paper3238 Reviewers Submitted, Paper3238 Authors

**Paper Summary:**

This paper proposes a novel self-query layer for self-supervised monocular depth estimation to capture the local pixel-level information. The proposed self-query layer consists of four parts; coarse-grained queries, building a self-cost volume, pixel-level attention approximation, and probabilistic combination to capture relative distance representations. The proposed model shows impressive results in recovering thin structures, and the experiments verify the effectiveness of the proposed method and generalization ability on popular datasets.

**Paper Strengths:**

* This paper is well written and easy to follow the high-level ideas. The authors present a good summary of almost all the major efforts in recent monocular depth estimation research based on scene consistency. The idea of building a self-cost volume to learn a representation for relative distance seems novel and makes sense. The coarse-grained queries improve computational efficiency and effectiveness (Table 5).
* Table 1 shows the effectiveness of the proposed method with different input resolutions on the popular KITTI dataset. The proposed method outperforms the previous works employing stereo supervision or multiple frames of input at test-time, which is a great success in the monocular self-supervision approach. The proposed method also shows great generalization performance in Table 4.
* The authors conduct ablation studies on every piece of their design choices, which verifies the effectiveness of the proposed method.

**Paper Weaknesses:**

[ Notations ]

* The size of the self-cost volume V must be Q x N (= Q x h/p x w/p) (L367). And if the S\_{j, k} in Eq. 4 is for the input of SQL layer (L319), S\_{j, k} must be a C-dimensional vector. Then how can we multiply the Q-dimensional vector (softmax(V\_i){j, k}\*) with the C-dimensional vector (S\*{j, k})? Also, what’s the size of the input tensor of the MLP module in pixel-level attention approximation?
* In Figure 3, where’s the pixel-wise softmax in every plane of V (L383)? where’s the plane-wise softmax (L407)?
* High-level idea is easy to understand, but the notations and equations could be revised and should be matched with Figure 3 for readers to understand the full details easily. I would suggest the authors revise Figure 3 to contain more detailed information.
* Please check and revise the notations and equations in Section 3.3.

[ Analysis of the efficiency ]

* It would be nice to have an analysis of the efficiency of the proposed model such as runtime / FLOPs, # of parameters, etc.

[ Inconsistency of the colormap ]

* The colormap inconsistency hinders comparison with other methods in Figure 1 and Figure 4. I believe this is due to the scale-ambiguity of the self-supervised methods, and the authors could adjust colors, e.g., normalizing predicted depth, to match the depth values.

[ Experiment on Cityscapes ]

* Even though the proposed method does not utilize the motion mask and semantic prior, initialization of the weights by the model trained on KITTI might not be a fair comparison since the other methods are just trained on Cityscapes.
* It would be great to see if the zero-shot transfer of the model trained on KITTI to Cityscapes achieves competitive performance. What if the SQLdepth model is just trained on Cityscapes from scratch (or ImageNet pre-trained weights)?

[ Typos ]

* L676: ECPDepth → EPCDepth [51]

**Overall Recommendation:**4: weak accept

**Justification For Recommendation And Suggestions For Rebuttal:**

* The current manuscript could be improved to help reproduce this idea in the future, but the proposed idea seems novel, and the effectiveness is shown in the experiments.
* Please check and revise the notations and equations in Section 3.3 and Figure 3.

**Confidence Level:**4: The reviewer is confident but not absolutely certain that the evaluation is correct.

**Final Rating:**2: weak reject

**Final Rating Justification:**

According to the rebuttal, the proposed SQL adds 9.68M parameters to the model, which is enormous considering the backbones used in the experiments have 27M or 25M parameters. The proposed method trained purely on CityScapes (yellow, AbsRel 0.137) shows worse performance than Monodepth2 (AbsRel 0.129), even though the proposed method has about 30% (27M +9.68M vs. 25M + a) more parameters. So the effectiveness of the proposed module is in doubt. Also, it is unclear if the performance gain in Table 1 is from changing backbone (EfficientNet-b5 vs. ResNet50).

The proposed idea sounds interesting, but its effectiveness is not clearly supported by experiments.

[[–]](https://openreview.net/forum?id=fF2UW7W4dR5" \o "Collapse reply thread)

## **Official Review of Paper3238 by Reviewer c46M**

*CVPR 2023 Conference Paper3238 Reviewer c46M*

06 Jan 2023 (modified: 08 Feb 2023)CVPR 2023 Conference Paper3238 Official ReviewReaders: Program Chairs, Paper3238 Senior Area Chairs, Paper3238 Area Chairs, Paper3238 Reviewers Submitted, Paper3238 Authors

**Paper Summary:**

This paper proposed a novel transformer-based output block for self-supervised depth estimation. This block uses the output features of an encoder-decoder CNN (EfficientNet-b5) to predict a depth cost volume with adaptive quantised depth bins. This new architecture is trained on KITTI Eigen and tested on KITTI Eigen, CityScapes (finetuned) and Make3D (zero-shot). The summary of claims is:

1. Proposes a new hybrid architecture, including an attention-based cost volume and adaptive discrete depth bins per-image.
2. SOTA performance on Kitti and CityScapes
3. Improved generalisation to new datasets (CityScapes and Make3D), showing adaptability to various depth distributions.

**Paper Strengths:**

1. The paper proposes an interesting self-supervised adaptation of AdaBins, which requires knowledge of the real ground-truth depth distribution.
2. Predicting depth as a probability distribution is also a sensible idea, as it more flexible than simple regression and potentially allows the model to capture the uncertainty in complex image-regions, such as depth boundaries.
3. The proposed model is testing on multiple datasets and seems to provide good performance in each of these.
4. The ablation for the proposed model is carried out well. Each contribution is tested out individually and—more importantly—also also incorporates multiple “common sense” baselines.

**Paper Weaknesses:**

1. ****Lack of attention to detail:**** The paper contains many typos, formatting and citation errors that distract from the content of the paper. This is not grounds for rejection, but I strongly encourage the authors to revise the paper per the comments below.
2. ****Lack of detail in literature review:**** Monocular training approaches are simply grouped as “dealing with moving scenes” and “extra constraints”. This does not provide the reader enough detail to understand prior literature. This review is also missing prior self-supervised work that utilises self-supervised cost volumes [A, B].
3. ****No discussion of computing costs:**** It would be interesting incorporate additional information about the model, such as the memory requirements or runtime performance, e.g. FLOPS, MParams, FPS…
4. ****The claims for improved generalisation are dubious:**** Models evaluated on CityScapes are fine-tuned on that dataset (which is larger than the base KITTI Eigen) and Make3D has low-quality ground-truth that results in unreliable evaluations. The paper also makes claims regarding generalising to different depth distributions. However, all tested datasets are outdoor urban scenarios. If the depth distributions are in fact significantly different, the paper should incorporate some figures demonstrating this. Maybe evaluating on the DDAD or DIODE datasets would help strengthen these claims, as these datasets have increased depth ranges up to 200 meters.
5. ****CityScapes evaluation:**** Are the other baselines evaluated on CityScapes fine-tuned on this dataset? If they are not, then this evolution is not really a fair comparison. If would also be valuable to show the performance of a zero-shot model without fine-tuning, which would strengthen the claims of generalisation.
6. ****Evaluation on KITTI Eigen:**** The main evaluation is carried out on KITTI Eigen, which has poor quality ground-truth and incorrect/saturated metrics. I would strongly recommend putting the results from the updated KITTI Eigen-Benchmark (currently in supplementary) in the main paper and leaving KITTI Eigen as supplementary. As a community, we should be moving away from these incorrect and unreliable benchmarks.
7. ****Non-comparable architectures:**** The proposed model uses a retrained Efficient-Net-b5 as its backbone, which most models use ResNet-18. Recent studies [C] have shown that changes in backbone architecture can account for most of improvement gains that are usually attributed to novel contributions. It would be valuable to show the performance of a base model trained using this common ResNet-18 backbone.
8. ****Incorrect claims about existing models:**** In the results, Monodepth2 is labeled as using a motion mask, which is incorrect. The text in 3.4. (masking strategy) makes is sound like most evaluated methods use a motion mask, which is also incorrect. For instance, PackNet, HR-Depth, Johnston, CADepth, DepthHints and DIFFNet do not use motion masks of any sort. Most of them use the minimum reproduction loss from Monodepth2, which is not the same.

[A] Juan Luis Gonzalez Bello and Munchurl Kim. Forget About the LiDAR: Self-Supervised Depth Estimators with MED Probability Volumes. In Advances in Neural Information Processing Systems, volume 33, pp. 12626–12637, 2020.

[B] Juan Luis Gonzalez Bello and Munchurl Kim. PLADE-Net:Towards Pixel-Level Accuracy for Self-Supervised Single-View Depth Estimation with Neural Positional Encoding and Distilled Matting Loss. In Conference on Computer Vision and Pattern Recognition, pp. 6847–6856, 2021. doi: 10.1109/CVPR46437.2021.00678.

[C] Jaime Spencer, Chris Russell, Simon Hadfield, Richard Bowden. "Deconstructing Self-Supervised Monocular Reconstruction: The Design Decisions that Matter". Transactions of Machine Learning Research 2022.

**Overall Recommendation:**4: weak accept

**Justification For Recommendation And Suggestions For Rebuttal:**

As commented in the “Strengths” section, the contributions proposed intuitively make sense and seem to provide significant improvements based on the ablation studies. However, the comparisons against existing methods and the claims for generalisation are not as strongly supported, as per the comments in the “Weaknesses” section. The rebuttal should address the following comments:

1. Comments in “Weaknesses”
2. Why are models using additional frames for prediction evaluated? These are not monocular models.
3. It is hard to tell if the visualisations are better. Why are the proposed method depth maps brighter, when the depth range is the same? Is a different normalisation scheme used? Authors should also consider using the “Turbo” colourmap [D], which enhances far depth details and is better for colourblind readers.
4. Incorporating point cloud visualisations (supplementary) is highly encouraged. However, they should be visualised from more extreme angles, as front views hide boundary interpolation artefacts and incorrect shapes.
5. Do the distributions learnt by the model reflect what we as humans would expect? E.g. depth boundaries might be represented as as 50/50 chance of being either foreground or background. Or does the model simply learn an arbitrary distribution of bins that, when doing the weighted sum, results in the correct depth?
6. Is there any intuition behind why increasing the number of queries beyond 128 does not further improve performance?

[D] Turbo, An Improved Rainbow Colormap for Visualization. Anton Mikhailov. (<https://ai.googleblog.com/2019/08/turbo-improved-rainbow-colormap-for.html>)

**Confidence Level:**4: The reviewer is confident but not absolutely certain that the evaluation is correct.

**Additional Comments For Authors:**

1. No paper ID was provided in the submission.
2. The SQL acronym should maybe be introduced earlier in the paper, e.g. even in the abstract.
3. What does the dagger symbol mean next to SQLdepth in all evaluation tables?
4. L93: dataset.. -> dataset.
5. L136: volumn -> volume
6. Fig 2: Wrap, wrapped -> warp, wrapped
7. Whole paper: CNN based, transformer based -> CNN-based, transformer-based
8. Orphaned table/figure/citation refs. Include a non-breaking space to prevent this. Table ??? -> Table~???.
9. L792: reception fields -> receptive fields
10. Why are most venues in the citations not capitalised? E.g. computer vision and pattern recognition
11. Duplicated references: [15] & [16], [29] & [30], [56] & [57], [82] & [83].
12. Please review all citations and ensure the details/venues are correct and consistent.
13. Why are the references in the supplementary material on the first page?

**Final Rating:**2: weak reject

**Final Rating Justification:**

Thank you to the authors for providing the rebuttal and the reviewers for the additional discussions. Unfortunately, the rebuttal does not really address my concerns and in fact seems to provide more confusion.

* The additional zero-shot results on CityScapes still do not address the lack of clarity regarding the settings for the baselines. What datasets were they trained on? Why are none of the "fair comparison" baselines from table 2 evaluated zero-shot on CityScapes? What does the yellow row represent in the zero-shot experiments? Is it comparable to Monodepth2 and why does it perform worse?
* The paper/rebuttal also keeps making reference to motion mask and semantic information. However, this does not apply to many of the baselines.
* The rebuttal does not address the issues regarding non-comparable backbones. There is still likely a performance gain from switching from ResNet to EfficientNet. The paper does not acknowledge this, meaning that some of the improvement gains are likely not due to the proposed contributions.
* I still do not agree that methods using temporal multi-frames are monocular. If a method uses more than one image to make a depth prediction, it is not monocular.