FREE-VIEW EXPRESSIVE TALKING HEAD VIDEO EDITING

Yuantian Huang, Satoshi Iizuka, Kazuhiro Fukui

University of Tsukuba

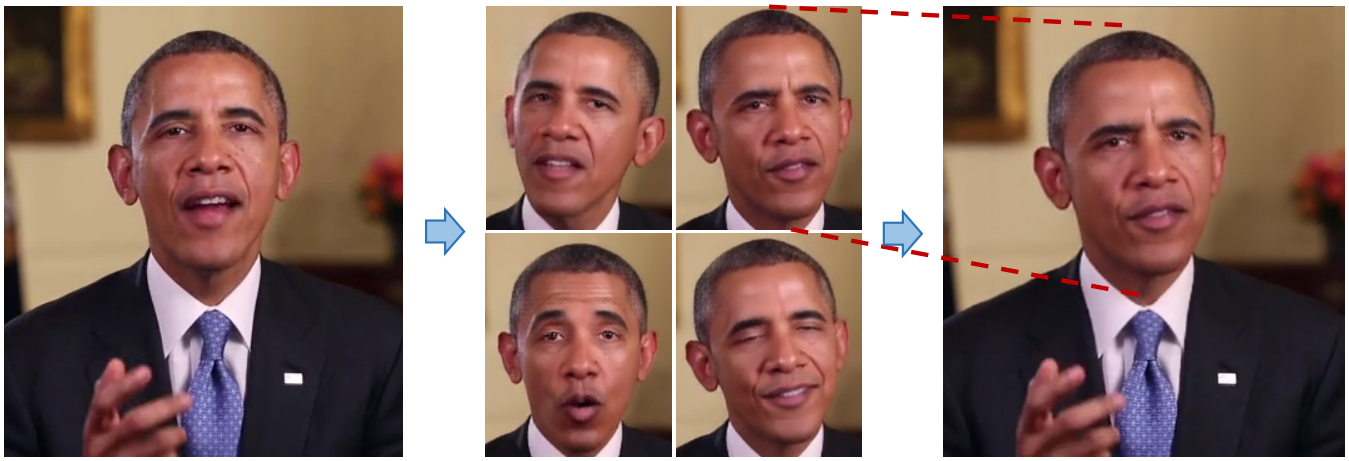

Fig. 1: Our approach can edit the talking head in a video with multiple freely controllable attributes, including head pose, facial emotion and eye blink. Edited frames can be restored to their full frames while maintaining audio-visual synchronization.

ABSTRACT

We present a novel framework for talking head video editing, allowing users to freely edit head pose, emotion, and eye blink while maintaining audio-visual synchronization. Unlike previous approaches that mainly focus on generating a talking head video, our proposed model is able to edit the talking heads of an input video and restore it to full frames, which supports a broader range of applications. Our proposed framework consists of two parts: a) a reconstruction-based generator that can generate talking heads fitting to the original frame while corresponding to freely controllable attributes, including head pose, emotion, and eye blink. b) a multiple-attribute discriminator that enforces attribute-visual synchronization. We additionally introduce attention modules and perceptual loss to improve the overall generation quality. We compare existing approaches as corroborated by quantitative metrics and qualitative comparisons.

Index Terms— Talking head, video editing.

1. INTRODUCTION

Audio-driven talking face generation has developed significantly in recent years, previous researchers[5, 6, 7, 8] tackle the problems of generating talking head videos from source audio. Although significant progress has been made in this area recently, most focus on improving lip synchronization while ignoring other vital factors related to overall naturalism, like head pose, facial expression, and eye blink. Some previous works keep the original head pose and eye blinks in a generated video, such as Wav2Lip [1]. Other works [9] are able to generate head movements inferred from an audio source, showing some degree of nature but not feasible to be freely controlled. More recently, PC-AVS [3] successfully applied head poses to its generated video by devising an implicit low-dimension pose code that is extracted by a reference video. EAMM [2] and GC-AVT [4] enables emotional talking face generation, which highly improves the quality and diversity of generated faces. However, these approaches only generate talking head videos from a source frame or a source video, while the generated video can not be restored to the full frame, limiting practical usage. Contradictorily, our proposed method can edit the talking head area and restore it back to the original video, leading to a much wider field of applications, such as video content editing, visual dubbing, and remote classroom and online lectures. Although other studies [10, 11] managed to edit talking heads in videos to some extent, they offer limited control over facial attributes and require complex 3D models for both training and inference.

In this paper, we design a reconstruction-based generator that can generate talking heads fitting the original frame. On the encoder side, the input frames that are masked on the head area are processed by a face encoder to embed face features. An audio encoder processes a Mel spectrogram ex-

Table 1: We compare the features of our proposed approach with different existing methods for talking head generation and editing. Specifically, Wav2Lip can edit in a full frame but lacks all the vital attributes for naturalism, while the other approaches failed to perform full frame editing. EAMM and GC-AVT can generate emotional talking head videos. Besides, PC-AVS and GC-AVT can apply head poses from reference videos, while our proposed method can adjust poses in free-view. Our approach is the most flexible of existing approaches.

Methods	Lip Sync	Full-Frame Editing	Head Pose	Emotion	Eye Blink
Wav2Lip [1]	✓	✓	Copy source		
EAMM [2]	✓		Copy source	✓	
PC-AVS [3]	✓		Reference required		
GC-AVT [4]	✓		Reference required	✓	
Ours	✓	✓	Free-view	✓	✓

tracted from the target audio, and additional encoders process head pose code, emotion code and eye blink code extracted from the target video into embedded features. These are then concatenated and fed to a decoder network with residual skip connections. In this pipeline, our proposed generator tries to in-painting the talking head out of masked frames while keeping audio and additional attributes synchronized to the video.

Moreover, we extend the idea of the lip-sync discriminator in Wav2Lip [1] into a multi-attribute discriminator that enforces not only the lip but also head pose and facial expression synchronization to the video. This enables a free-view expressive talking head generation for our proposed model, which significantly improves the accuracy and quality of our generated results.

Our proposed framework allows users to synthesize talking head videos with multiple attributes, such as head pose, emotion, and eye blinks. More importantly, our method enables restoring the generated talking head back to where it cropped, enabling editing videos in any size and scale. In most scenes, we assume users prefer to see a talking person, including hand and upper body movements, instead of a talking head alone. A high-level comparison of similar approaches is summarized in Table 1.

In summary, in this work we present:

- A reconstruction-based generator that can generate talking heads fitting to the original frame while corresponding to freely controllable attributes, including head pose, facial emotion, and eye blink.
- A multi-attribute discriminator that enforces attribute-visual synchronization.
- By combining these two modules, our proposed model can edit talking head videos in full frames on multiple attributes during the synthesis with or without references, which is one of the earliest attempts at this task.

2. PROPOSED FRAMEWORK

Our model consists of three sub-components: a) encoders that process multiple inputs into embedding features. b) a generator that generates frames corresponding to the input. c) a pre-trained multi-attribute discriminator shared a similar structure to encoders that can enforce attribute-visual synchronization. An overview of the model is shown in Fig. 2.

2.1. Model Input

The input of our proposed model are as followings:

- **Input frames** are five source frames and five target frames with the masked head area. The image size is 256×256
- **Audio** input is **Mel spectrograms** calculated by corresponding audio to the target frames.
- **Head pose code** is a $b \times 3 \times t$ tensor that represents Euler angles (yaw, pitch, and roll), which are predicted by a pre-trained Hopenet [12].
- **Emotion code** is a $b \times 7 \times t$ tensor that represents seven emotion categories, calculated from a pre-trained RMN [13].
- **Eye blink code** is a $b \times 1 \times t$ tensor that represents eye openness, computed from face landmarks using dlib [14].

Here, b is the batch size, t is the time step in which we set it to five. In the inference stage, the target frames could be input frames themselves, and other attributes can be manually defined or extracted from reference frames.

2.2. Model Architecture

Encoders. We use five different encoders to process different input data with a similar structure, a stack of 2D convolutions, as shown in the left side in Fig. 2. All the outputs are the same as $b \times D \times t$, which are then concatenated together and fed to the generator.

Generator. We use a similar generator as Wav2Lip, borrowed initially from LipGAN ([15]), which is a stack of 2D convolutions with transpose convolutions for up-sampling. Additionally, we adopt dual attention modules [16] into our generator to improve image generation quality. We train the generator using L_1 loss and perceptual loss [17] that are computed between generated frames and target frames. We also adopt sync losses from Multi-Attribute Discriminator. we use a blink loss that tries to improve the blink accuracy, calculated by eye landmarks distance between the generated frames and target frames.

Multi-Attribute Discriminator. Based on the observation that Wav2Lip [1] has a significant improvement in lip synchronization compared to other existing approaches, we assume a pre-trained discriminator is highly efficient for syn-

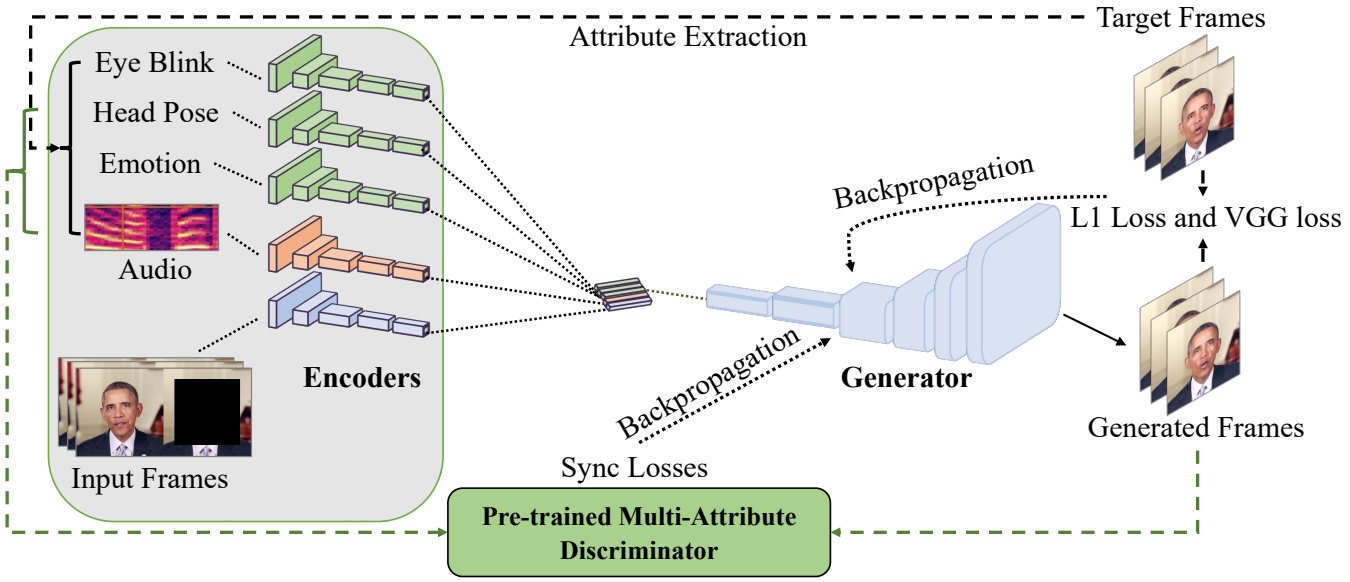

Fig. 2: Structure of our framework. The encoders embed inputs together and feed them into the generator, while the input audio Mel spectrogram, head pose, emotion, and eye blink are extracted from target frames during the training stage. A set of synchronization losses are then calculated by a pre-trained multi-attribute discriminator between generated frames and input attributes to enforce attribute-visual synchronization.

chronization tasks. Therefore we extend the lip-sync discriminator from Wav2Lip into a more powerful multi-attribute discriminator, which not only processes audio and visual data but also process additional features, including head pose code and facial emotion code. We minimize the L_2 distance between embedding features that are generated from the face encoder and those from other encoders.

$$L_{sync} = L_{audio} + L_{pose} + L_{emotion} \quad (1)$$

Our full objective is formulated as follows:

$$L_{total} = L_{l1} + \lambda_v L_{perceptual} + \lambda_s L_{sync} + \lambda_b L_{blink} \quad (2)$$

3. RESULTS

3.1. Datasets

We use 5 hours of youtube videos of Barack Obama, initially collected by the Synthesizing Obama [18]. Besides that, we manually collect 5 hours of President Joe Biden from youtube and borrow part of the dataset proposed by Merkel Podcast Corpus [19], which is also around 5 hours. We use 90% of each dataset as a train set and 10% for validation. Currently, our model is trained on a single person for higher accuracy.

3.2. Experiments

We choose two leading models in talking head generations. Wav2Lip [1] can edit in full frames but is not capable of applying additional attributes, while PC-AVS [3] can change head poses to its generated video but can not restore it to full frames. We train Wav2Lip on the same datasets as our proposed model. Regarding the PC-AVS model, we could not train it on our datasets as the training code of PC-AVS has not been publicly released. Therefore, we used the pre-trained model on VoxCeleb2 [20] provided by the authors instead.

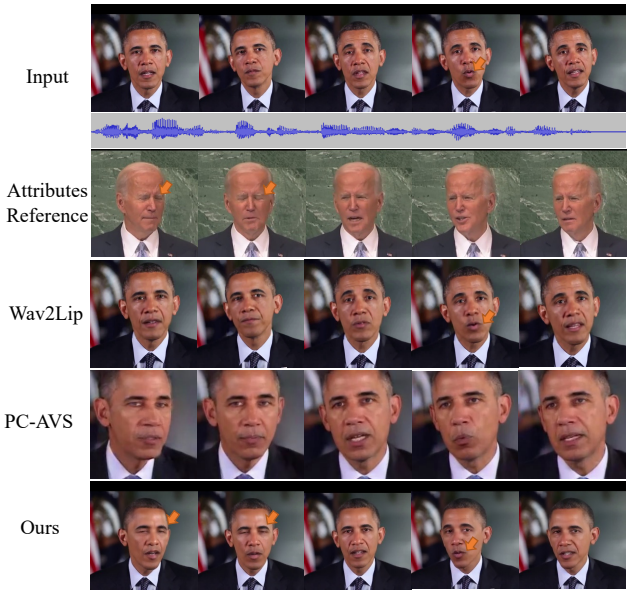

Fig. 3: Visual Quality Comparisons against existing approaches. Note that PC-AVS can only generate faces based on one frame, while the Wav2Lip model and our proposed model can restore the generated faces to the original frames.

We set our problem as editing attributes in an input video, that is, take attributes reference from another video, and apply head pose, emotion, and blink to the original video while keeping its synchronization. We use the validation dataset as described in Sec. 3.1. Specifically, we randomly select 100

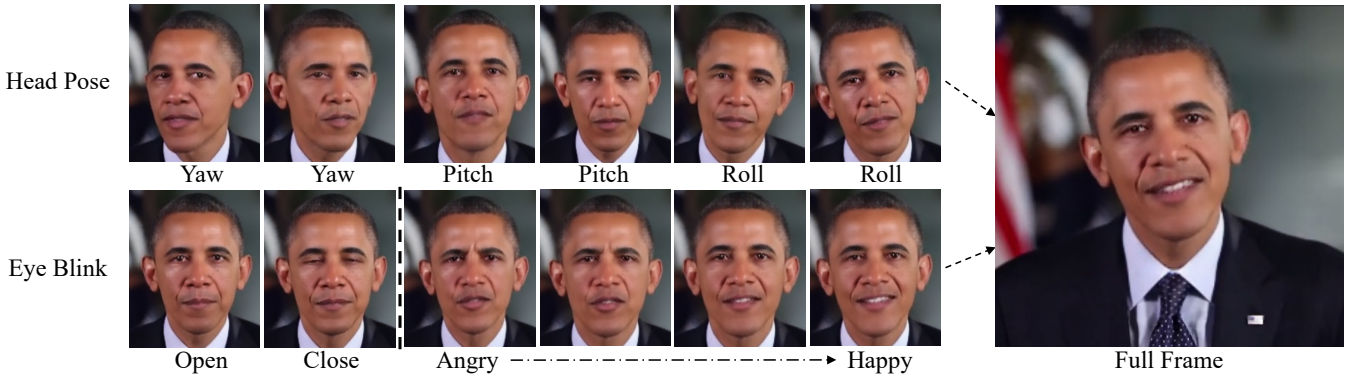

Fig. 4: Examples of Editing different types of attributes and restoring to the original frame.

Table 2: Quantitative comparison with existing approaches using metrics.

Methods	SSIM	LMD_m	LMD_f	SyncNet
Wav2Lip [1]	0.71	0.57	1.39	**5.67**
PC-AVS [3]	0.68	0.63	3.16	4.39
Ground Truth	1.00	0.00	0.00	5.24
Ours	**0.72**	**0.46**	**0.86**	4.79

pairs of short videos in each dataset. For each pair, a video will be used as input frames and audio, the other one is used for attribute referencing.

Note that EAMM [2] model and GC-AVT [4] model does not publicly release their code for evaluation since they have been published very recently. Thus we are not able to compare our results with these two methods. However, as discussed in Table 1, EAMM can only apply emotion and GC-AVT can not change eye blinks, and both methods are not capable of editing in full frames.

3.3. Quantitative Comparison

We use **SSIM** [21] to evaluate the video generation quality, **SyncNet** [22] scores and $\mathbf{LMD}_m$ [23] to evaluate mouth movements, $\mathbf{LMD}_f$ to evaluate how well the expression, eye blink fits the desired results. As shown in Table 2, the result shows that our proposed model outperforms baseline models in terms of generation quality and landmark accuracy. Note that Wav2Lip is trained under the same SyncNet discriminator, resulting in the highest SyncNet confidence scores. However, our model is comparable in lip synchronization and more flexible in many other aspects.

3.4. Visual Quality Comparison

We provide a qualitative comparison of results in Fig. 3 Obviously, Wav2Lip fails to apply any changes since it only accepts audio input, while PC-AVS can change head poses properly. However, it is a one-shot generation model, which can not give these changes back to the original frames. Baseline models also fail to apply emotion and eye blink, as pointed out by the arrow.

Table 3: Ablation study using the FID metric. We evaluate the effect of each element, which is described as head Pose(P), Emotion(E), and eye Blink(B). The (A + P + E + B) is our full model.

Methods	SSIM	LMD_m	LMD_f	SyncNet
A(Only Audio)	0.70	**0.45**	1.13	**4.88**
A + P	0.70	0.49	0.95	4.73
A + P + E	0.69	**0.46**	0.90	4.78
A + P + E + B	**0.72**	**0.46**	**0.86**	4.79

3.5. Attribute Editing

Our proposed framework is able to fully control three different types of facial attributes, including head pose, facial expression, and eye blink. An example of Editing different facial attributes is shown in Fig. 4.

3.6. Ablation Study

We also conduct an ablation study to verify the effectiveness of our multi-attribute discriminator. Each attribute is described as head Pose(P), Emotion(E), and eye Blink(B). The results are shown in 3, showing that performance increases when corresponding elements are added.

4. CONCLUSION

We have presented a novel framework that is able to edit specific attributes of a talking head video while maintaining audio-visual synchronization. Our proposed framework is one of the earliest challengers at this goal, and evaluations demonstrate that our model outperforms existing models. The main limitation of our proposed model is that it may not be able to generate satisfactory results when training with multiple identities. We expect a general model that includes identity learning in our future research.

5. REFERENCES

[1] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 484–492.

[2] Xinya Ji, Hang Zhou, Kaisiyuan Wang, and et al. Wu, "Eamm: One-shot emotional talking face via audio-based emotion-aware motion model," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, SIGGRAPH '22.

[3] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4176–4186.

[4] Borong Liang, Yan Pan, and et al. Guo, "Expressive talking head generation with granular audio-visual control," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 3377–3386.

[5] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *European conference on computer vision*. Springer, 2020, pp. 716–731.

[6] Yuanxun Lu, Jinxiang Chai, and Xun Cao, "Live speech portraits: real-time photorealistic talking-head animation," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 6, pp. 1–17, 2021.

[7] Yudong Guo, Keyu Chen, and et al. Liang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5784–5794.

[8] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3661–3670.

[9] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman, "You said that?: Synthesising talking faces from audio," *International Journal of Computer Vision*, vol. 127, no. 11, pp. 1767–1779, 2019.

[10] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala, "Text-based editing of talking-head video," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–14, 2019.

[11] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.

[12] Nataniel Ruiz, Eunji Chong, and James M Rehg, "Fine-grained head pose estimation without keypoints," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 2074–2083.

[13] Pham Luan, Vu Huynh, and Tran Tuan Anh, "Facial expression recognition using residual masking network," in *IEEE 25th International Conference on Pattern Recognition*, 2020, pp. 4513–4519.

[14] Davis E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[15] Prajwal K R, Rudrabha Mukhopadhyay, and et al. Philip, "Towards automatic face-to-face translation," in *Proceedings of the 27th ACM International Conference on Multimedia*. 2019, p. 1428–1436, Association for Computing Machinery.

[16] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[18] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman, "Synthesizing obama: Learning lip sync from audio," jul 2017.

[19] Debjoy Saha, Shravan Nayak, and Timo Baumann, "Merkel podcast corpus: A multimodal dataset compiled from 16 years of angela merkel's weekly video podcasts," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France, June 2022, European Language Resources Association.

[20] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[21] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[22] Joon Son Chung and Andrew Zisserman, "Lip reading in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 87–103.

[23] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu, "Lip movements generation at a glance," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 520–535.